Soccer is one of the most popular team sports all over the world. Most sports games are naturally organized into successive and alternating plays of offense and defense, cummulating at events such as goals or attacks. Soccer video matches always attract a major sports audience. Recently, the amount of digitized video content has been increasing rapidly, and watching a soccer match needs a lot of time. Many sports fans prefer to watch a summary of soccer video matches. The summarization process is an essential part of several applications such as Information retrieval, video retrieval, etc. Soccer video summarization and analysis is concerned with the extraction of valuable semantics by efficient and effective processing of visual, audio, and text information.

Hossam M. Zawbaa

Lecturer, Information Technology Dept., Faculty of Computers and Information, Beni Suef University, Egypt. Dr. Zawbaa has published several papers in major international journals and peer-reviewed conference proceedings. He has a wide research scope that includes, Computer Vision, Image Processing, Machine Learning, Data Mining, and Biometrics.

978-3-659-35918-7

Automatic Soccer Video Summarization

Zawbaa

Hossam M. Zawbaa

Automatic Soccer Video Summarization

LAP LAMBERT Academic Publishing

Hossam M. Zawbaa

Automatic Soccer Video Summarization

Hossam M. Zawbaa

Automatic Soccer Video Summarization

LAP LAMBERT Academic Publishing

Impressum / Imprint
Bibliografische Information der Deutschen Nationalbibliothek: Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Bibliographic information published by the Deutsche Nationalbibliothek: The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Coverbild / Cover image: www.ingimage.com

Verlag / Publisher:
LAP LAMBERT Academic Publishing
ist ein Imprint der / is a trademark of
AV Akademikerverlag GmbH & Co. KG
Heinrich-Böcking-Str. 6-8, 66121 Saarbrücken, Deutschland / Germany
Email: info@lap-publishing.com

Herstellung: siehe letzte Seite /
Printed at: see last page
ISBN: 978-3-659-35918-7

Zugl. / Approved by: Cairo, Cairo University, Diss., 2012

Cairo University
Faculty of Computers and Information
Information Technology Department

Automatic Soccer Video Summarization

By
Hossam Mohammed Zawbaa Ismail

A Thesis Submitted to
Faculty of Computers and Information
Cairo University

In partial fulfillment of the
Degree of Master in Information Technology

Under the supervision of

Prof. Dr. Aboul Ella Hassanien

Dr. Nashwa El-Bendary

Dr. Iman Atef El-Azab

Faculty of Computers and Information
Cairo University
Egypt
2012

Abstract

Soccer is one of the most popular team sports all over the world. Most sports games are naturally organized into successive and alternating plays of offence and defence, cummulating at events such as goal or attack. If a sport videos can be segmented according to these semantically meaningful events, it then can be used in numerous applications to enhance their values and enrich the user's viewing experiences. According to this, soccer video summarization and analysis has recently attracted much research and a wide spectrum of possible applications have been considered.

Soccer video summarization and analysis is concerned with the extraction of valuable semantics by efficient and effective processing of combination of visual, audio and text information. However, one of the major limitations of current soccer analysis is the semantic gap between the low level features such as (color, texture, shape and motion) and high level representation such as (shot types, shot length, and shot replays).

This thesis presents an automatic soccer video summarization system using machine learning (ML) techniques. The proposed system is composed of five phases. Namely; in the pre-processing phase, the system segments the whole video stream into small video shots. Then, in the shot processing phase, it applies two types of classification (shot type classification and play / break classification) to the video shots resulted from the pre-processing phase. Afterwards, in the replay detection phase, the proposed system applies two machine learning algorithms, namely; support vector machine (SVM) and artificial neural network (ANN), for emphasizing important segments with championship logo appearance. Also, in the excitement event

detection phase, the proposed system uses both machine learning algorithms for detecting the scoreboard which contain an information about the score of the game. The proposed system also uses k-means algorithm and Hough line transform for detecting vertical goal posts and Gabor filter for detecting goal net. Finally, in the event detection and summarization phase, the proposed system highlights the most important events during the match. Experiments on real soccer videos demonstrate encouraging results. The event detection and summarization has attained recall 94% and precision 97.3% for soccer match videos from five international soccer championships.

To my parents, all my family members and my supervisors who gave their love, support, and time freely.

Acknowledgements

At first, at last and all the time, thanks to ALLAH the God of the world, for every thing in my life. Nothing in my life could be done without his permission, and no success could be gained without his mercy. Thanks to Prof. Aboul Ellah Hassanien for his very much support and encouragement to accomplish this thesis in a professional and valuable way. Although being very busy everyday, Prof. Aboul Ella Hassanien opened a door for research work in front of me and continually advised and encouraged me. Without his help, I would never finish my research work and this thesis. His rigorous work attitude, wide knowledge and sagacious vision have influenced me deeply and will benefit me continually throughout my career.

I would like to express my sincerest gratitude to my supervisors, Dr. Nashwa El-Bendary, and Dr. Iman Atef El-Azab, for their invaluable guidance and suggestions during my two years of study. It is wonderful to study under their supervision.

Finally, I would like to thank my parents, my sister and brothers, and, my friends for their constant support and encouragement. I will love you forever.

List of Publications

Journal Papers:

1. **Hossam M. Zawbaa**, Nashwa El-Bendary, Aboul Ella Hassanien "Automatic Soccer Video Summarization System," Submitted to Signal, Image and Video Processing (SIViP), (Springer journals), 2011.

2. **Hossam M. Zawbaa**, Nashwa El-Bendary, Aboul Ella Hassanien, and Tai-hoon Kim, "Event Detection Based Approach for Soccer Video Summarization Using Machine learning," Submitted to Science & Engineering Research Support society (SERSC), (Indexed by SCOPUS, EI), 2011.

Peer Reviewed International Conference:

3. **Hossam M. Zawbaa**, Nashwa El-Bendary, Aboul Ella Hassanien, and Tai-hoon Kim, "Machine Learning-based Soccer Video Summarization System," International Conference on Multimedia, Computer Graphics and Broadcasting (MulGraB 2011), CCIS/LNCS series Springer, (Indexed by SCOPUS, EI) December 8 ~ 10, 2011 in Jeju Grand Hotel, Jeju Island, Korea , pp. 19-28, 2011.

4. **Hossam M. Zawbaa**, Nashwa El-Bendary, Aboul Ella Hassanien, Gerald Schaefer and Sang-Soo Yeo, "Support Vector Machine based Logo Detection in Broadcast Soccer Videos," The 16th Online World Conference on Soft Computing in Industrial Applications (WSC16) Advances in Intelligent and Soft Computing, 5-6 Dec., Springer 2011.

5. **Hossam M. Zawbaa**, Nashwa El-Bendary, Aboul Ella Hassanien, and Ajith Abraham, "SVM-based Soccer Video Summarization System," The Third IEEE World Congress on Nature and Biologically Inspired Computing (NaBIC 2011) Salamanca University, Spain during October 19-21, pp. 7-11, 2011.

Contents

List of Figures

List of Tables

List Of Abbreviations

ML	Machine learning techniques
SVM	Support vector machine classifier
ANN	Artificial neural network classifier
RGB	Red green blue color space
HSI	Hue saturation intensity color space
ART	Adaptive resonance theory
BP	Back propagation neural network
RBP	Revised back propagation neural network
RBF	Radial basis function neural network classifier
HT	Hough Transform
CAF	Confederation of African football championship
AVI	Audio video interleave

LIST OF TABLES

Chapter 1

Introduction

This chapter presents the importance of soccer video analysis and summarization. Soccer video matches always attract major sports audience. Recently, the amount of digitized video content has been increasing rapidly and watching a soccer match needs a lot of time, many TV fans of sport competitions prefer to watch a summary of soccer video matches. A summary is presented about the problems that are facing the automatic soccer video summarization and the proposed solutions. Finally an overview about the organization of the thesis is shown at the end of the chapter.

1.1 Background

Fully automated video summarization represent a hotly pursued research topic in the field of content-based video analysis. Most sport games are naturally organized into successive and alternating plays of offence and defence, cumulating at events such as goal or attack. If a sport video can be segmented according to these semantically meaningful events, it then can be used in numerous applications to enhance their values and enrich the user's viewing experiences.

Soccer video matches always attract major sports audience. Recently, the amount of digitized video content has been increasing rapidly and users need to access their content through various network facilities. As watching a soccer match needs a lot of time, many TV fans of sport competitions prefer to watch a summary of football games (1). According to this, soccer video analysis has recently attracted much research and

a wide spectrum of possible applications have been considered. Traditionally soccer videos are analyzed manually but this costs valuable time. Therefore it is necessary to have a tool that does the job automatically.

Soccer video analysis is concerned with the extraction of valuable semantics by efficient and effective processing of combination of visual, audio and text information. However, one of the major limitations of current soccer analysis is the semantic gap between the low level features such as (color, texture, shape and motion) and high level representation such as (shot types, shot length and replays).

This thesis proposes such a solution that targets at bridging the semantic gap and building an innovative intermediate representation of high level and low level video information to aid in indexing, retrieval, and browsing. This solution is based on an understanding of broadcast soccer video.

1.2 Motivation

Nowadays, with the progress in video compression, storage and communication, we are able to put a large amount of digital videos in database or online for users to perform query for some interesting or meaningful data. While the amount of video data is rapidly increasing, multimedia applications are still very limited in content management capability. Therefore, mining information in video data becomes an increasingly important problem as digital video becomes more and more pervasive.

The ubiquitous consumption of video, however, poses many problems among which the field of multimedia processing focuses on the effective description of video information (video modeling), the relationship between low level features and semantic meanings of video information (video processing / analysis), and the querying of such information for fast and easy access to the relevant set at a later time (video querying / video search and retrieval).

Automatic video indexing becomes one of the major challenges in the field of information systems. The automatic soccer video summarization extracts the important events to produce general summaries for the most important moments in which soccer viewers may be interested. Researchers have proposed many techniques to take full advantage of the fact that sport videos have typical and predictable temporal structures, recurrent events, consistent features, and fixed number of camera views.

The multimedia analysis tool, which could automatically parse soccer video and output required video clips or the most interesting events such as goals, corner kicks and free kicks, fans could go though many more games without spending much time. This can entertain these fans and in turn further popularize the sport itself. So, soccer video indexing, especially event detection is absolutely necessary.

Event detection in soccer video is a high level analysis, which needs an effective description of soccer video information and approaches to bridge the gap between low level features and semantic meanings as its foundations. However, research in this field is far from enough. Shot is commonly used as an intermediate representation, but its propriety for soccer video parsing needs to be further studied and other high level representations should be explored. This thesis work has been inspired by this motivation.

We have two propellants which motivated us to develop an automated system for soccer match summarization. First, most people cannot watch the whole matches which are played on same time within different time zones because of lack of time. Second, coaches need to view the highlighted events to truly developing plans and evaluate the team players. From this point we concluded the importance of our proposed program to put a solution for the mentioned problems.

1.3 Problem statement

Analyzing general sport games is still an open problem because of the variance and diversity of different games. Some former researchers have proposed many highlight summarization methods both for general sports game and for a specific kind of sports

game (2). detected the play and break event in sports videos to generate the summary. Some other researchers summarize sports videos using slow motion replays (3). On the other hand, another group of researchers turn to study specific sport games such as soccer, basketball or diving (4).

Summarization process is an essential part in several applications such as (Information retrieval, video retrieval, etc), to retrieve the important parts. This field is undergoing rapid change, as computers are now prevalent in virtually every application, from games for children through the most sophisticated planning tools for governments and multinational firms.

When we are talking about soccer game, we can refer to a continuous sports which mean that if there is an existence of such a break during the match, it can an indicator of the occurrence of important event such as (goal, penalty shot and red / yellow card). Therefore the summarization process which we aim for can be recognized by a combination of these events, so the summarized segment may contain only the goal shots, goal attempts or penalty shots that can be described as important events.

The input is a soccer video match needed to be summarized using a computer based application, our concern here is to extract the most exciting events in the soccer game such as (goal, attacks, and other events) using our proposed application then output those events into summarized video.

In this thesis, we are going to highlight the most important events such as (goals, attacks, and the other events), to facilitate the process of automatic match, save the viewer's time, and introduce the technology of computer based summarization into sports field.

Soccer video analysis and summarization is concerned with the extraction of valuable semantics by efficient and effective processing of combination of visual, audio and text information. However, one of the major limitations of current soccer video summarization is the semantic gap between the low level features and mid level representation.

1.4 Overview of the mid level Representation

This section proposes such a solution that targets at bridging the semantic gap and building an innovative intermediate representation of high level and low level video information to aid in indexing, retrieval, and browsing. This solution is based on an understanding of broadcast soccer video.

The goal of this research work is to define and realize an appropriate mid level representation for soccer video analysis. In order to achieve a reliable video description, the primary requirement is to structure the video into elementary shots. This video partitioning step enables us to provide content based browsing of the video.

Based on our study of the soccer video structure, the soccer video shots can be classified into categories associated with semantic descriptors (close-up, audience, global, medium, play, and break). So, a soccer video stream can be divided into segments, each of which belongs to one or more of these categories. In another word, this video sequence can be delineated by a corresponding semantic descriptor sequence. This sequence is the proposed mid level representation, which facilitate higher level tasks such as video editing or retrieval.

In order to convert a soccer video stream into a semantic descriptor sequence, a computational framework is proposed and two approaches are devised to realize this framework. There are two phases, pre-processing phase, and shot processing phase. To reduce the computational complexity, motion difference is used to preliminarily segment a soccer video stream into relatively small parts (shots) in the pre-processing stage. After that each shot can be classified into categories (close-up, audience, global, medium, play, and break) in the shot processing phase.

We use replay detection to efficiently localize the highlights. Then incorporating with domain specific knowledge, we adopt several significant cues to evaluate the importance degree of the highlights and classifying it into goals, attacks, and other events.

1.5 Scope of work and proposed solutions

Proposed solutions have been presented to avoid most of the problems discussed in the previous section, these solutions are involved in each of the stages of the proposed system . In the preprocessing stage, grass color extraction and shot boundary detection (5) is proposed to the system segments the whole input video stream into small video shots. Also shot processing phase, it applies two types of classification (shot type classification and play/break classification) to the video shots resulted from the pre-processing phase. Afterwards, in the replay detection phase, the system applies two machine learning algorithms (6) , namely; support vector machine (SVM) and artificial neural network (ANN), for emphasizing important segments with logo appearance.

Subsequently, in the excitement event detection phase (6, 7) the proposed system detect some cinematic features; like (scoreboard, goal mouth, etc), for scoreboard detection, the system uses both ML algorithms for detecting the caption region providing information about the score of the game. The system uses k-means algorithm and hough line transform for detecting vertical goal posts and gabor filter for detecting goal net. Finally, in the logo based event detection and summarization phase, the system highlights the most important events during the match.

1.6 Thesis Organization

The thesis consists of six chapters including this introduction chapter. The introduction explores the characteristics of the input soccer video, the importance of the automatic soccer video summarization and the problems that are facing the automatic soccer video summarization and the proposed solutions.

Chapter 2 shows the discussed previous work on video segmentation, retrieval and those related closely to soccer video analysis and summarization are reviewed.

Chapter 3 introduces the support vector machine based for automatic logo detection in broadcast soccer videos. This chapter composed of three phases as follows:

- **Pre-processing phase** that segments the whole video stream into small video shots.
- **Shot processing phase** that applies two types of classification to the video shots resulted from the pre-processing phase.
- **Replay detection phase** we applies two machine learning algorithms, namely; support vector machine (SVM) and artificial neural network (ANN), for emphasizing important segments with logo appearance.

Chapter 4 shows an event detection and summarization, where we focus on the other addition phases of our proposed system. This chapter composed of two phases as follows:

- **Excitement event detection phase** the proposed system applies two machine learning algorithms, namely; support vector machine (SVM) and neural network (NN), to detect the scoreboard which contains an information about the score of the game. The proposed system uses k-means algorithm and Hough line transform for detecting vertical goal posts and Gabor filter for detecting goal net.
- **Event detection and summarization phase** the proposed system highlights the most important events during the match.

Chapter 5 shows the experimental work performed and the characteristics of the data used in testing the proposed approaches or techniques. This chapter includes the empirical evaluation of the proposed techniques where the results for each technique are discussed. It also includes a comparison between the different approaches and the proposed techniques.

Finally, chapter 6 summarizes the results obtained from the empirical investigation presented in the chapter of experimental work. The impact of these results is discussed with respect to the proposed techniques. Finally, a number of issues that have been raised by this work are then discussed, and presented as directions for future work.

Chapter 2

Related work

2.1 Introduction

Multimedia information systems are increasingly important with the advantage of broadband networks, high powered workstations, and compression standards. Compared with still images, videos are dynamic data with the temporal dimensions. That means a video cannot be only regarded as a sequence of still images with information in temporal dimensions ignored. While lots of techniques are developed in image retrieval, unique features of video data give rise to many new challenging issues.

The purpose of this thesis is to discuss semantic soccer video summarization, so the theory and methods used in soccer video summarization need to be carefully studied. In this chapter, existing works on video segmentation and retrieval are surveyed in the first and second sections because it can help us understand commonly used approaches in video analysis. With these understandings, we can better study related work in soccer video summarization, which is discussed and compared in the third section, and finally we get overview about the machine learning techniques in the last section.

2.2 Video segmentation

Video structure parsing is an initial step to organize the content of videos. Video data are typically organized in a typical hierarchical structure as shown in Figure 2.1 In this step, some elementary units such as scenes, shots, frames, key frame and objects are

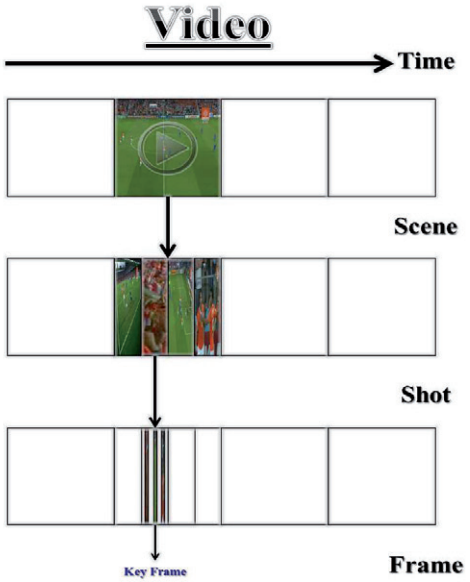


Figure 2.1: Hierarchical structure of video

generated. A successful structure parsing is important for video indexing, classification and retrieval. In the past, many works have been done in video structure parsing, especially in shot detection, motion analysis and video segmentation.

As discussed above, video data are structured into many shot units. Shot changes should be detected before dividing video data into shot units. A shot change can be viewed as detection of a camera break. Normally, there are three major editing types of camera breaks: cut, wipe and dissolve. A cut is an immediate change from a shot to another shot; a wipe is a change where first frame of a shot is replaced with last frame of another shot gradually; a dissolve is a change where one shot gradually appears (fade-in) and another shot slowly disappears (fade-out). A cut can be detected

by comparing two adjacent frames. While wipe and dissolve are difficult to detect since they are change gradually. The transition between shots usually corresponds to a change of subject, scene, camera angle, or view. Therefore, it is very natural to use shots as the unit for video indexing and analysis.

There are many works for detection of camera breaks in the past few years. They can be grouped into two categories: uncompressed and compressed domain. Some typical methods for the detection of camera breaks could be found in (8)(9)(10)(11). Recent published papers for shot change detection could be found in (12)(13)(14)(15)(16)(17)(18). Most work has been focusing on pixel difference, intensity statistics comparison, histogram distance, edge difference, and motion information. Among these methods, histogram based ones have been consistently reliable, while DCT coefficient based ones give the lowest precision. Motion information based methods are somewhere in between. Some work for performance evaluation of shot detection could be found in (19)(20).

Some work has been done on detecting these special effects. Related works can be found in (21)(22)(23)(24). In a recent review paper, Lienhart (17) compares four major shot boundary detection algorithms, which include fade and dissolve detection. Extensive experimental results also favor the color histogram based method (25) for shot boundary detection, instead of the computationally expensive edge change ratio method (26).

2.3 Video retrieval

Video segmentation is not a goal in itself but just a means for further analysis. For example, it can be used in video retrieval. We have already looked at work in video segmentation; from now on, related work in video retrieval will be surveyed. To date, most video retrieval systems are used to retrieve similar video based on low level features. Video retrieval faces the same problem with image retrieval that it lacks a semantic model and effective representation tool to express human perception.

There exists a gap between high semantic concept and low level features. How to bridge the gap is the most challenging topic in video classification and retrieval research. In this section, we will survey recent work on similarity based retrieval, clustering based video retrieval and semantic video retrieval.

2.3.1 Similarity based video retrieval

In current video retrieval system, there are two methods used for retrieval: similarity based and cluster based methods (9). Similarity based method is employed to retrieve similar video key frame, shot or video scene segment. Similarity matching can be based on the features extracted locally or globally. In a simple way, similarity measure is based on computing the similarity of related key-frame between two videos.

More sophisticated methods are employed the spatio-temporal features of video frames between two videos (27)(28)(29). Dagtas et al. (27) presented several motion descriptors as intermediate motion model for event based video retrieval. They retrieved the event videos by computing the similarity of different motion models. Chang et al. (28) proposed a method to retrieve video object by computing similarity of motion trajectories and trails in the spatial and temporal domains. Chang et al. also presented a semantic visual template, which can express the semantic concept (30). Detailed explanation of the idea will be discussed in later section.

2.3.2 Clustering based video retrieval

Clustering method is introduced as a solution to organize the content of video collections. It provides efficient method to classify and index the video since similar videos are clustered into similar group. Recent work on cluster based retrieval can be found in (31)(32)(33). In (31), Clarkson et al. proposed a framework to find the event by clustering the nature input audio/visual data. They developed a system that can cluster the video data into events such as passing through doors and crossing the street. The clustered events can also be clustered into high level scene.

2.3.3 Semantic video retrieval

Semantic video retrieval of video content is viewed as the promising trend of computer vision and multimedia. Effective semantic retrieval of video is a way to ultimate multimedia understanding. Currently most works are focusing on frame based structure modeling. Fully automatic multimedia understanding is almost impossible in state of the art. Although it is a very challenging work, there still exist some good research work resided on this topic (30)(31)(32)(33)(34)(35)(36)(37)(38).

In (37), Naphade et al. proposed a probabilistic framework for modeling multimedia object called Multiject and modeling semantic concepts called Multinet. Multiject can represent low level feature, such as visual features, audio features and textual features. It can also express the intermediate level meaning such as semantic template [54] and other high level semantic concepts. The advantages of a multinet is that it provides a framework for support four aspects of semantic indexes. One of its disadvantages is that the complexity of the framework will increase exponentially when the scope of knowledge is increased.

In (38), Chang et al. provide a Semantic Visual Templates (SVT) for modeling the low level feature and high level semantic object. They introduced the idea of SVT to bridge the gap between the users information needs and what the systems can deliver. Although the semantic visual template can express the semantic concept intuitively, however it can only describe some basic and simple semantic concept. It is quite difficult to represent a high level semantic event concept by sketching an intuitive template. Previously , many research had been proposed to extract and abstract the video objects in order to model the semantic concepts of objects and events.

In (39), Hwang et al. proposed a scheme for object based abstraction and analysis and semantic event modeling. However, based on the state of the art in computer vision, it is difficult to build such a system since the semantic features modeling depends on domain specific knowledge.

2.4 Soccer video summarization

As the most popular sport, soccer game attracts billions of people. However, even the most faithful fans cannot finish hundreds of games taken on a weekly basis. So, video indexing, especially event detection in soccer videos is absolutely necessary. Methods used in video segmentation and retrieval have already been reviewed above. As a genre of video, soccer video can be analyzed by these methods with some modification. In this section, some important research related to soccer video summarization are reviewed and compared. This is helpful in having clear idea about what have been done and what need to be further studied.

Y.H. Gong et al. in (40) proposed a system that can automatically parse soccer video programs using domain knowledge. The parsing process was mainly built upon line mark recognition and motion detection. They categorized the position of the play into several predefined classes by recognizing the compound line pattern with signature method. The motion vectors field is used to infer the play positions for those scenes without line marks. Despite the strong semantic indexes from the categorization of play positions, they have to address these two problems:

- How to identify different camera angle and shooting scale, otherwise the line mark recognition cannot be robust.
- How to determine reasonable segment for processing.

D. Yow et al. in (41) presents techniques to automatically detect and extract the soccer highlights by analyzing the image contents, and to present these shots of action by the panoramic reconstruction of selected events. The analysis include the recognition of prominent features of the game, tracking of ball, camera movement compensation for effective recognition, and construction of the panoramic views. The authors pointed out a direction for application of soccer video summarization.

V. Tovinkere et al. in (42) present an effective data mining framework for automatic extraction of goal events in soccer videos. The extracted goal events can be used for high level indexing and selective browsing of soccer videos. The proposed multimedia data mining framework first analyzes the soccer videos by using joint multimedia

features (visual and audio features). Then the data pre-filtering step is performed on raw video features with aid of domain knowledge, and the pre-filtered data are used as the input data in the data mining process using classification rules. The proposed framework fully exploits the rich semantic information contained in visual and audio features for soccer video data, and incorporates the data mining process for effective detection of soccer goal events. This framework has been tested using soccer videos with different styles as produced by different broadcasters. The results are promising and can provide a good basis for analyzing the high level structure of video content.

O. Utsumi et al. in (43) proposed a novel object detecting and tracking method in order to detect and track objects necessary to describe contents of a soccer game. On the contrary to intensity oriented conventional object detection methods, the proposed method refers to color rarity and local edge property, and integrally evaluate them by a fuzzy function to achieve better detection quality. These image features were chosen considering the characteristics of soccer video images, that most non-object regions are roughly single colored (green) and most objects tend to have locally strong edges. We also propose a simple object tracking method, which could track objects with occlusion with other objects using a color based template matching. The result of an evaluation experiment applied to actual soccer video showed very high detection rate in detecting player regions without occlusion, and promising ability for regions with occlusion.

P. Xu et al. in (44) introduced a framework for play / break events detection in soccer video. In this paper, three kinds of views in soccer video, global, zoom-in and close-up, are predefined. The counterpart's terms of these views are long shot, medium shot, and close-up, respectively. Here the grass value and classification rules are learned and automatically adjusted to each new clip. Then heuristic rules are used in processing the view label sequence, and obtain play / break status of the game. The system is novel, but it is just a good start for further event detection in soccer video.

A. Ekin et al. in (2) presented a fully automatic and computationally efficient framework for analysis and summarization of soccer videos using cinematic and object based features. In this paper, algorithms of dominant color region detection, robust

shot boundary detection and shot classification, as well as goal detection, referee detection, and penalty box detection are discussed. The algorithm of dominant color region detection is very impressive, but the methods used in goal detection and referee detection depend heavily on man made rules. Three types of summaries can be automatically produced:

- All slow motion segments in a game.
- All goals in a game.
- Slow motion segments classified according to object based features.

L.Y. Duan et al. in (45) presented a unified framework for semantic shot classification in sports videos. Unlike previous approaches, which focus on clustering by aggregating shots with similar low level features, the proposed scheme makes use of domain knowledge of a specific sport to perform a top down video shot classification, including identification of video shot classes for each sport, and supervised learning and classification of the given sports video with low level and middle level features extracted from the sports video. This framework looks good but still has some problems:

- Where the test data came from is not clearly mentioned.
- Methods used to detect flying graphics are too specific.
- Methods for shot classification is mainly based on shot segmentation, which is done by some commercial software.

Other works such as (46)(47) are also related to soccer video summarization. With consideration of our research work, a comparison among (2)(44)(45) is given in Table 2.1.

Paper	Function of System	Feature used	Classes of Shot	Result	Contribution	Comment
P. Xu	View classification Grass Orientation Classification Play/Break Segmentation	Color (hue) Texture Rules	Long Shot Medium Shot Close-up	For view classification, 85% For Play / Break Segmentation, 75%	Color-based grass detector Play/Break Segmentation	The thresholds for different games in view classification are different Using simple rules to do Play/Break Segmentation
L.Y. Duan	Shot Classification	Color Motion Texture Camera motion pattern Dominant Object Motion Homogeneous Regions Rules	Close-up Field View, Following, Player Medium Still, Audience, Corner Kick, Goal View, Replay	85% -95%	Relationship between shot and semantic meanings Mapping from low-level features to mid-level features Mid-level features representation Fusion of valid mid-level features at shot level Real-time Performance	The method for Flying Graphics detection is too specific Method for detection of Field-Players Interaction Curve (FPIC) may be heavily affected by bad light conditions Where their testing data came from is not clearly mentioned
A. Ekin	Shot classification Slow-motion detection Event detection	Color Motion Rules	Long Shot In-field media shot Close-up / Out-field shot Goal Referee Penalty Box	For cut detection, 97.3% recall and 91.7% precision. For Gradual transitions, 85.3% recall and 86.6% precision. short classification 88%.	Field color detection Novel features for shot classification	Make full use of color information But haven't put much effort on how to use motion features

Table 2.1: Comparison of research work in soccer video summarization (48)

Chapter 3

Machine learning based logo replay detection

3.1 Introduction

This chapter introduces the concept of dominant color detection, shot boundary detection, different classes of shots (Long, Medium, Close-up, and Audience (out of field)) based on the camera position, and the video editing techniques used to produce replay shots and the used methods to detect them. The automatic logo detection from soccer videos proposed in this paper is composed of the following three fundamental building phases:

- **Pre-processing phase** that segments the whole video stream into small video shots,
- **Shot processing phase** that applies two types of classification to the video shots resulted from the pre-processing phase,
- **Replay detection phase** that applies support vector machine (SVM) and artificial neural network (ANN) for emphasizing important segments with a logo appearance.

These three phases are described in details in the following sections along with the steps involved and the characteristics feature for each phase, see the figure 3.1.

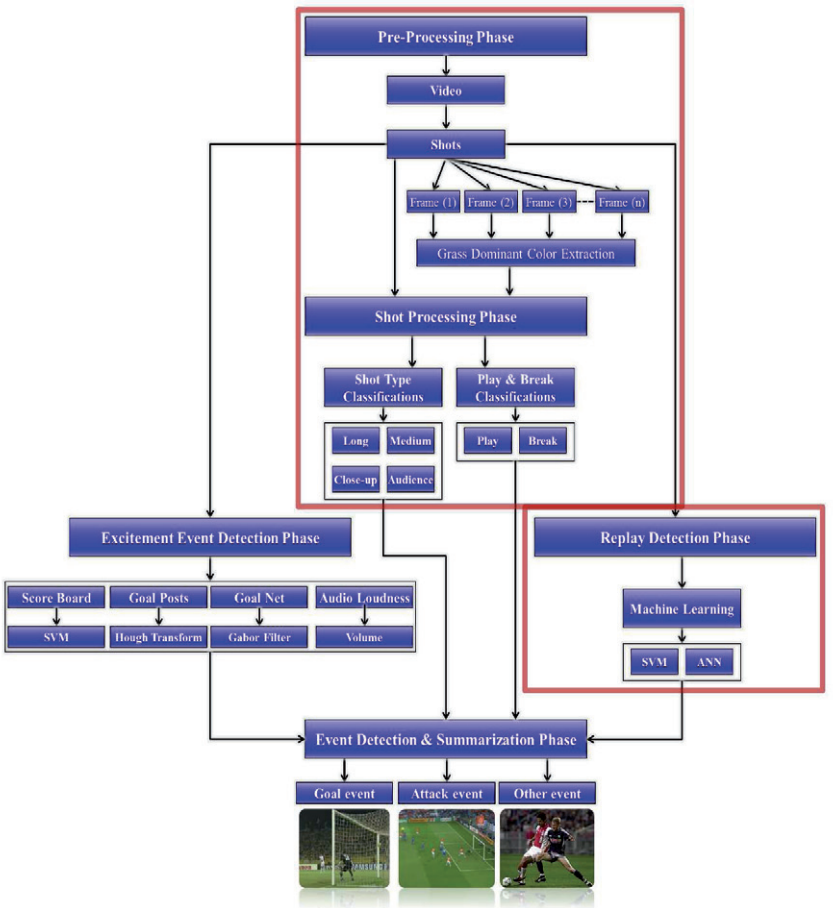


Figure 3.1: System architecture - logo replay detection phases

3.2 Pre-processing phase

The goal of this phase is to segment the whole video stream into small video shots. By detecting the dominant color in the video frame, then the shot boundary detection algorithm is applied in order to output video shots based on dominant color derived features.

3.2.1 Dominant color detection

The dominant color is the color that fills most of the given area, and it is different among various play fields. For example, in figure 3.2 we can clearly see that the dominant color differs from one sport to another. In this thesis, we are concerning only with the soccer game, which has a green color for the playing field.

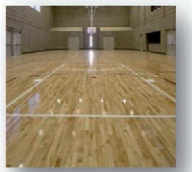
The playing field usually has a distinct tone of green that may vary from stadium to another stadium. But in the same stadium, this green color may also change due to the weather and lighting conditions, as you see in Figure 3.3. As dominant color extraction is challenging due to effects on the play field such as shadow, lighting, low resolution and other environmental factors, there are several color spaces that have been used for the dominant color detection including HSI and RGB.

3.2.2 Color spaces

Color segmentation algorithms have high importance to extract various features that depend on the selected color space and must be done truly with fewer errors. For this reason we mention the color spaces types and their conversion functions which are used in our system and discuss their usage in the implemented methods (1).

Color space is defined as a model for representing color in terms of intensity values. For most image acquisition devices, the video format of the video stream determines the color space of the acquired image data, that is, the way color information is represented numerically. For example, many devices represent colors as RGB values. In this color space, colors are represented as a combination of various intensities of red, green,

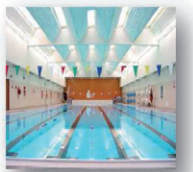
(a) Soccer



(b) Basketball field

(c) Hockey field

(d) Tennis Court



(e) Swimming pool

(f) Football field

Figure 3.2: Different dominant color of the sport fields

and blue. The most common color space used is RGB. Another color space, The HSI color space (Hue, Saturation, Intensity) is often used by people who are selecting colors (e.g., of paints or inks) from a color wheel or palette, because it corresponds better to how people experience color than the RGB color space does. RGB space reduces light effect, regardless of the current weather or stadium condition, or even the quality of the video. In the other hand, HSI doesn't provide the mentioned facilities (12).

As hue varies from 0 to 1.0, the corresponding colors vary from red through yellow, green, cyan, blue, magenta, and back to red, so that there are actually red values both at 0 and 1.0. As saturation varies from 0 to 1.0, the corresponding colors (hues) vary from unsaturated (shades of gray) to fully saturated (no white component). As intensity, or brightness, varies from 0 to 1.0, the corresponding colors become increasingly brighter. The following figure 3.4 illustrates the HSI color space.

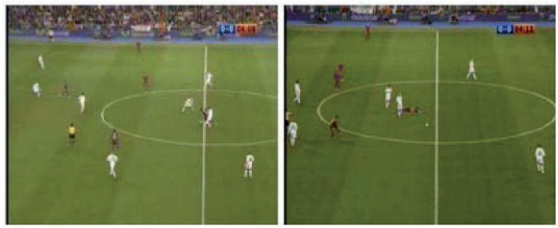


Figure 3.3: Different lighting conditions of soccer fields

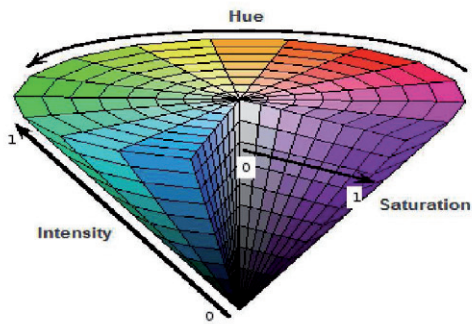


Figure 3.4: HSI color space

Hue family of color spaces, such as HSI and HSV; decouple intensity component from the chrominance information that is represented by hue and saturation values. In addition to that, hue and saturation closely match human perception. Although there are multiple definitions, we use equations for the conversion of RGB to HSI: Since hue is unstable for low saturation values, hue cannot be reliably used for certain range of color values. This range of color values is referred to as the achromatic region in the HSI space (2).

Most approaches today suffer from factors which have several negative effects on the image quality such as (lighting, shadows, low resolution etc). You can't ignore this

Figure 3.5: Dominant color extraction: The first row represents the original images and the second row represents dominant color in white (binary) images

problem if your goal depends on the quality that is affected by the degree of visibility of relevant information in an image which is appropriate for a specific image task such as (detection, classification etc).

3.2.3 Grass dominant color extraction

Distinguishing field colors from others is not as easy as one may think because the RGB values may change under different lighting and field conditions or different camera shooting positions. Authors in (2) used a self adapted method to detect field color in HSI color space. It is effective but too complex. P. Xu et al. in (44) detect field color With consideration of accuracy and complexity. We designed a method to solve this problem as in algorithm (1). Algorithm (1) shows the steps for grass dominant color extraction. Also, the dominant color extraction examples are shown in figure 3.5.

3.2.4 Shot boundary detection

Video processing and computer vision communities usually employ shot based structural video models, and associate low level descriptors such as color, texture, shape and motion, and semantic descriptions in the form of textual annotations, with these structural elements. But, there are very little work that aims to bridge the gap between the low level features and semantic descriptions to arrive at a well integrated structural

Algorithm 1 Grass Dominant Color Extraction

1: Convert the input video file into its corresponding frames
2: **for** each frame **do**
3: Convert the frame from RGB to HSI color space using equations (3.1), (3.2), and (3.3)

$$H = \cos^{-1}[\frac{\frac{1}{2}[(R-G)+(R-B)]}{[(R-G)^2+(R-B)+(G-B)]^{\frac{1}{2}}}] \tag{3.1}$$

$$S = 1 - \frac{3}{R+G+B} \times \min(R+G+B) \tag{3.2}$$

$$I = \frac{1}{3}(R+G+B) \tag{3.3}$$

4: Define the color range that covers the different variations of the play-field's green color
5: **for** Each pixel **do**
6: **if** The three HSI components with range of values: $(0.15 < H < 0.4)$, $(0.1 < S < 1)$, $(0.2 < I < 1)$ **then**
7: Set the color of this pixel to "$White$"
8: Otherwise; set the color of this pixel to "$Black$"
9: **end if**
10: **end for**
11: **end for**

semantic video model.

Based on our study of soccer video structure, we found that for the purpose of semantic description, shot is not suitable as a mid level representation (e.g. too long to be delineated by a semantic word). This means video analysis on a shot basis could not fully use all the essential information contained in soccer videos, which will result in the limitation in further analysis such as event detection and summarization. Instead, we introduce a structural semantic video representation for efficient description of low level video features.

Firstly, we define some categories for soccer video classification, and semantic de-

scriptors (long, medium, close-up, audience, play, and break) shot views are associated with them to delineate their semantic meanings. Therefore, a soccer video stream can be divided into segments, each of which belongs to one or more of these categories. Or, this video stream can be delineated by a semantic descriptor sequence. This is our proposed mid level representation of the soccer game. Thus, a soccer video stream can be represented by a sequence of descriptors after segmentation and classification processing (2).

Soccer video summarization and analysis are concerned with the extraction of valuable semantics by efficient and effective processing of combination of visual, audio and text information. However, one of the major limitations of current soccer analysis is the semantic gap between the low level features and mid level representation. The next sections propose such a solution that targets at bridging the semantic gap and building an innovative intermediate representation of high and low level video information to aid in indexing, retrieval, and browsing. This solution is based on an understanding of broadcast soccer video.

3.2.4.1 Shot based representation drawbacks

Traditionally, structural video summarization represents video as a union of smaller coherent shots that are obtained by a temporal or a spatio-temporal segmentation process. The boundaries of these temporal shots correspond to large differences in some feature space while a temporal shot has similar features within itself. These features are usually a combination of color, texture, shape, and motion, which are commonly referred to as low level features.

A shot can be defined as a collection of frames recorded during a continuous motion of the camera. There are two main reasons of doing this:

- Simplify computational complexity in video processing.
- The assumption of shots in a video stream can be regarded as a natural segmentation.

Figure 3.6: Instant (cut) shot transition

Hence, the frames within a shot represent a continuous action in time and space, and share the same high level features as well as similar low level features. Thus, the frame to frame similarity within a shot is exploited to generate compact video representations by key frames, which refer to one or more frames in a shot that best represent its content (46).

3.2.4.2 Transitions in soccer video stream

According to (49), there are three major types of camera breaks: cut, wipe and dissolve. A camera cut is an instantaneous change from one shot to another; a wipe is a moving boundary line crossing the screen such that one shot gradually replaces another; a dissolve superimposes two shots where one shot gradually lighten while the other fade out slowly. Wipe and dissolve are normally referred to as gradual transitions. as shown in figure 3.6 and figure 3.7, respectively.

According to statistic data in (49), more than 70% of all kinds of transitions are cut and less than 30% are other kinds of transitions; a sports video clip almost always contains both cuts and gradual transitions. So, the detection of these kinds of transitions except for cut should be more important if we insist to do shot segmentation.

Figure 3.7: Gradual shot transition

3.2.4.3 Shot boundary detection algorithm

The separated views that come from multiple cameras positioned at different locations. It can be realized that while changing from one camera to another, this indicates a start of a new shot and marks a boundary of a new shot. Accordingly, a shot can be defined as sequence of frames recorded by a single camera with a continuous action in time and space (2). Algorithm (2) shows the steps of shot boundary detection.

3.3 Shot processing phase

In order to convert a soccer video stream into a semantic descriptor sequence. Each shot can be classified into categories (close-up, audience, long, medium, play, and break) in the shot processing phase. This phase applies two types of classification; namely, shot type classification and play / break classification, to the video shots resulted from the pre-processing phase (5).

3.3.1 Shot type classification

Production crews use different shot types to convey make different scenes, which can be used for high level video analysis in a particular domain. Cinematographers classify a shot into one of four categories long, medium, close-up and audience (out of field) shot classes as presented in figures 3.8 (a), (b), (c), and (d), respectively., the definitions of which are usually domain dependent. In the following, we define these four classes for sports video:

Algorithm 2 Shot Boundary Detection

1: **for** each frame **do**
2: Convert the frame from RGB to HSI color space
3: Do frame skipping by step (k = 10 frames) to convert a gradual transition into a cut transition
4: Calculate the motion difference= the difference between current hue frame and the next hue k-frame
5: Divide the original color frame into 32 x 32 blocks size
6: Compute the mean change = the average blocks change between the current frame and the next k-frame
7: Compute the total percentage of changed blocks = the total percentage of changing blocks between the current frame and the next k-frame
8: Compute the grass ratio = the average difference of grass dominate color between the current frame and the next k-frame
9: **if** (Mean change $>$ Thr) and (total percentage of changing blocks $>$ 0.25) and (grass ratio $>$ 0.1) **then**
10: Mark a new shot
11: **end if**
12: **end for**

- **Long shot** displays the global view of the field; long shots almost always display some part of the stadium, which decreases the dominant colored pixel ratio. The long shot serves for an accurate localization of the events on the field.

- **Medium shot** where a whole human body is usually visible, is a zoomed-in view of a specific part of the field.

- **Close-up shot** usually shows the above-waist view of a player or referee.

- **Audience shot** The audience, coach, and other shots are denoted as out of field shots.

The sequence occurrence of a close-up shots and audience (out of field) indicates an important event such as (goal, goal attempts etc) during the match (2)(12).

(a) (b) (c) (d)

Figure 3.8: Different shot classes

The average dominant color ratio was used for all frames during the shot for determining the view type of that shot. The shot type classification algorithm proposed in this paper is based on a specific threshold (range) for grass ratio (G). It has been developed offline by training each class with different shots from different matches in order to define a specific range for each one of the four shot types based on dominant color ratio histogram and assign each shot to one of the four different shot types.

A set of thresholds have been defined for distinguishing the grass ratio for the different shot types (2). For the proposed system, we applied four threshold ratios, each frame can be classified into one of the previously stated views depending on equation (3.4). Where, GL stands for grass ratio of long shot view, GM stands for grass ratio of medium shot view, GC/A stands for grass ratio of both close-up and audience (out of field) shot views. Furthermore, additional thresholds were used for distinguishing between close-up and audience shot views. BA was used as the average black color ratio in the whole frames during shot.

$$Shot-type = \begin{cases} Long-View, & G_L \geq 0.55; \\ Medium-View, & G_M \geq 0.15 \text{ and } G_M < 0.55; \\ Close-up-View, & G_{C/A} \leq 0.15 \text{ and } B_A < 0.8; \\ Audience-View, & G_{C/A} \leq 0.15 \text{ and } B_A \geq 0.8. \end{cases} \tag{3.4}$$

3.3.2 Play / break classification

The sports video is a repetitions of play and break scenes. Some researchers claim that play sequences in sports videos are self consumable because most users naturally focus

their attention on events that happen within plays. A play scene is generic because it can contain a sequence of shots where the ball is being played in a soccer match. Generally, successive long and medium view shots usually correspond to play event, while frequent shots and/or successive close-up and out of field view shots indicate break events that cause a game to stop momentarily (such as a foul, celebrating a goal, or the end of a playing period). For example, when a whistle is blown during a play, users would expect that something happened. During the break, the close-up views of the players and/or a replay scene will confirm whether it was a foul or offside. Unlike previous work that categorizes sports videos into either highlights or play sequences, we aim to present a unifying summarization framework that integrates highlights into plays as well as reveal why we should still retain breaks (50)(5).

Using the start and end frames location of view type classified shots, the boundaries of each play and break shot can be determined. The start of a play shot is identified as the first frame of consecutive long view frames, which can be interleaved by very short zoom-in or close-up shots. On the other hand, the start of a break shot is identified as the starting frame of either a long zoom-in shot or a shorter close-up shot, which can be interleaved by very short long view shots. Consecutive play shots are considered as a play scene, which usually are ended with a consecutive break shots. Thus, a play/break sequence is a combination of consecutive play and break scenes, and sport games consist of many of this sequences (51). The play/break actions are described in figure 3.9.

3.4 Replay detection phase

A Replay is a good semantic cue to detect events due to frequently happening after every event in sports games, especially in the soccer domain. Therefore, we use Replay as one basic event in our method. The replay shot is detected by taking into account the fact that before and after replay shot, there usually is a very short shot of a flying logo of a certain broadcaster or an organization (52).

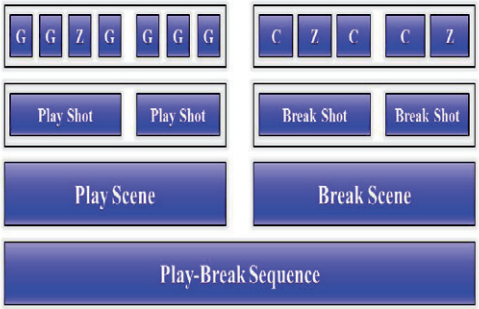


Figure 3.9: Play/break classification

In most soccer matches, exciting events are often replayed. These exciting shots normally correspond to highlights in a game, e.g., actions near the goal posts in a soccer game. Replay is a video editing way that is often used to emphasize an important segment with a logo appearance for once or several times. The inputs to replay detection phase are the output shots from the shot boundary detection step of the pre-processing phase in order to extract the exciting events that are represented by the replay shots. In sports video, there is often a highlighted logo that appears at the start and end of a replay segment, which indicates an exciting event within the soccer match. In the recent years, broadcasters use inserted logo sequence, as a digital video effect to replay the exciting and important events in soccer videos (1), as presented in figure 3.10 and figure 3.11 that shows an examples of gradual logo appearance (5).

3.4.1 Machine learning techniques

Machine learning techniques (ML) play an important role to solve many complex problems especially, in the logo replay detection and score board detection. In this section we present an overview of the ML techniques.

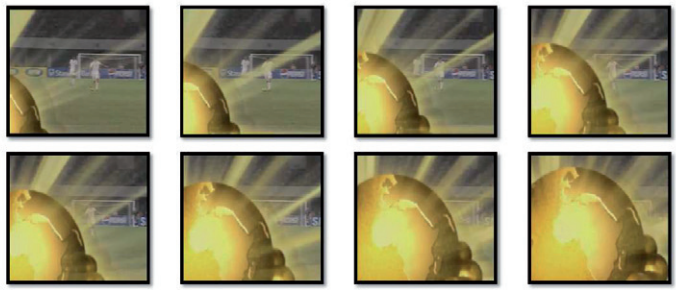


Figure 3.10: Gradual logo appearance of CAF 2008

Figure 3.11: Gradual logo appearance of CAF 2010

3.4.1.1 Support vector machine (SVM)

Kernel based techniques (such as SVM, Bayes point machines, kernel principal component analysis, and Gaussian processes) represent a major development in ML algorithms. SVMs were first suggested by Vapnik in the 1960s for classification and have recently become an area of intense research owing to developments in the techniques and theory coupled with extensions to regression and density estimation (53). SVM are a group of supervised learning methods that can be applied to classification or regression. Classification is achieved by a linear or nonlinear separating surface in the input space of the dataset. SVM delivers state of the art performance in real world applications such as text categorization, handwritten character recognition, image classification, bio-sequences analysis, etc., and is now established as one of the standard tools for ML and data mining. It employs kernel to map the input data into some much higher dimensional feature space implicitly in which data becomes linear separable. The linear decision boundary is drawn in a manner that the margin, minimum distance between training examples to the boundary, is maximized. In case that the mapped data points are linearly inseparable, a cost is included to account for the wrongly classified examples and the margin is maximized together with minimizing the cost.

The SVM approach seeks to find the optimal separating hyperplane between classes by focusing on the training cases that are placed at the edge of the class descriptors. These training cases are called support vectors. Training cases other than support vectors are discarded. This way, not only is an optimal hyperplane fitted, but also less training samples are effectively used; thus high classification accuracy is achieved with small training sets (54).

According to the idea, the spam filtering can be viewed as the simple possible SVM classification application of linearly separable classes; that is, a new email either spam or safe email. A complete formulation of SVM can be found at a number of publications (54). Here, the basic principles will be presented. The SVM algorithm seeks to maximize the margin around a hyperplane that separates a positive class from a negative class.

Given a training dataset with n samples $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, where x_i is a feature vector in a v-dimensional feature space and with labels $y_i \in -1, 1$ belonging to either of two linearly separable classes C_1 and C_2. Geometrically, the SVM modeling algorithm finds an optimal hyperplane with the maximal margin to separate two classes, which requires to solve the optimization problem, see equations 3.5 and 3.6:

$$maximize \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j . K(x_i, x_j) \tag{3.5}$$

$$Subjectto : \sum_{i=1}^{n} \alpha_i y_i, 0 \leq \alpha_i \leq C \tag{3.6}$$

where α_i is the weight assigned to the training sample x_i. If $\alpha_i > 0$, x_i is called a support vector. C is a regulation parameter used to trade-off the training accuracy and the model complexity so that a superior generalization capability can be achieved. K is a kernel function, which is used to measure the similarity between two samples. Different choices of kernel functions have been proposed and extensively used in the past and the most popular are the gaussian RBF, polynomial of a given degree, and multi layer perceptron (53). These kernels are in general used, independently of the problem, for both discrete and continuous data.

The SVM is a theoretically superior ML methodology with great results in classification of high dimensional datasets and has been found competitive with the best ML algorithms. More researchers pay attention to SVM based classifier for spam filtering, since their demonstrated robustness and ability to handle large feature spaces makes them particularly attractive for this work.

Islam et.al. in (55) proposed an innovative and intelligent spam filtering model based on SVM. This model combines both linear and nonlinear SVM techniques where linear SVM performs better for text based spam classification that share similar characteristics. The proposed model considers both text and image based email messages for classification by selecting an appropriate kernel function for information transformation.

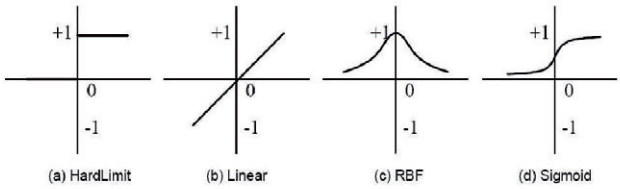


Figure 3.12: Basic transfer functions

Dual v-Support Vector Machines (2v-SVM) addresses the difficulties that arise when the class frequencies in training data do not accurately reflect the true prior probabilities of the classes. Authors in (55) applied 2v-SVM over Chinese email volume, which improves the classification precision of waste email management system.

3.4.1.2 Artificial neural network (ANN)

Artificial neural networks have been developed as generalizations of mathematical models of biological nervous systems. In a simplified mathematical model of the neuron, the effects of the synapses are represented by connection weights that modulate the effect of the associated input signals, and the nonlinear characteristic exhibited by neurons is represented by a transfer function. There are a range of transfer functions developed to process the weighted and biased inputs, among which four basic transfer functions widely adopted for multimedia processing are illustrated in Figure 3.12.

The neuron impulse is then computed as the weighted sum of the input signals, transformed by the transfer function. The learning capability of an artificial neuron is achieved by adjusting the weights in accordance to the chosen learning algorithm. Most applications of neural networks fall into the following categories:

- **Prediction** Use input values to predict some output.
- **Classification** Use input values to determine the classification.

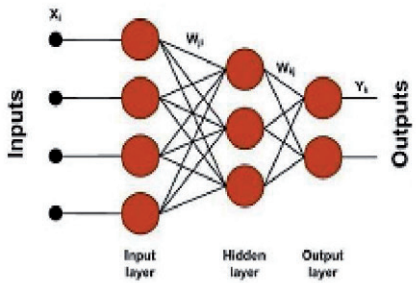


Figure 3.13: Hierarchical structure of video

- **Data Association** Like classification but it also recognizes data that contains errors
- **Data conceptualization** Analyze the inputs so that grouping relationships can be inferred.

The behavior of the neural network depends largely on the interaction between the different neurons. The basic architecture consists of three types of neuron layers: input, hidden and output layers.

In feed-forward networks the signal flow is from input to output units strictly in a feed-forward direction. The data processing can extend over multiple units, but no feedback connections are present, that is, connections extending from outputs of units to inputs in the same layer or previous layers. There are several other neural network architectures (Elman network, adaptive resonance theory maps, competitive networks etc.) depending on the properties and requirement of the application. Reader may refer to (56) for an extensive overview of the different neural network architectures and learning algorithms. see figure 3.13.

Adaptive resonance theory (ART) was initially introduced by Grossberg (57) as a theory of human information processing. ART neural networks are extensively used for supervised and unsupervised classification tasks and function approximation. There are

many different variations of ART networks available today (58). For example, ART-1 performs unsupervised learning for binary input patterns, ART-2 is modified to handle both analog and binary input patterns, while ART-3 performs parallel searches of distributed recognition codes in a multilevel network hierarchy. ART-MAP combines two ART modules to perform supervised learning while fuzzy ART-MAP represents a synthesis of elements from neural networks, expert systems, and fuzzy logic.

Neural networks can be used to solve the problem of classifying emails into emails that internet users wish to receive and spam emails that internet users do not wish to receive. The main difficulty in the application of neural network for spam filtering is the high dimensionality of the input feature space. This is because each unique term in the vocabulary represents one dimension in the feature space, so that the size of the input of the neural network depends upon the number of stemmed words. The selected features should balance the tradeoff between the classification accuracy and efficiency (59). As well as, it requires considerable time for parameter selection and network training.

The conventional back propagation (BP) neural network has slow learning speed and is prone to trap into a local minimum, so it will lead to poor performance and efficiency. Also, a well constructed thesaurus has been recognized as a valuable tool in the effective operation of text classification, it can also overcome the problems in keyword based spam filters which ignore the relationship between words. Bo and Zhu in (59) proposed two spam filtering methods using revised back propagation (RBP) neural network and automatic thesaurus construction. The RBP neural network is used to overcome the limitations of the conventional BP neural network. Experimental results show that the proposed spam filtering system is able to achieve higher performance, especially for the combination of RBP neural network and automatic thesaurus construction.

3.4.2 Logo based replay detection

The aim of logo detection is the identification of a wide variety of logos used by different broadcasters. Based on experimental findings, a generic logo model can be developed

as:

- **Logo** is meant to be contrasted from the background.
- **Usually animated** within 10-20 frames with a general pattern of *smallest-biggest-smallest.*
- **Biggest contrast** usually takes at least 40-50 % of the whole frame, whereas smallest contrast is up to 10-20% (60).

For logo detection module, there are several processes passed by the input shot to determine the boundary of the replay scene, which is bounded by the logo. There is one feature, which is *Logo is appeared at the beginning of the shot*, generated by the shot detection stage used in the logo detection module. That feature is used to enhance the logo detection performance as well as its accuracy because only few frames from the beginning of each shot will be processed. Algorithm (3) describes the steps of Logo detection algorithm (61).

Algorithm 3 Logo Detection Using SVM and ANN Classifiers

```
Train the ANN classifier with correct logo and false logo samples
Train the SVM classifier with correct logo and false logo samples
for Each frame do
    Adjust image intensity values for increasing the contrast of the input frame
    Select region of interest based on color for returning a binary image
    Calculate frame white ratio = the percentage of white pixels in the whole frame
    if Any frame contains a large contrast object (the white frame ratio be greater
    than 0.5) then
        Get the original colored frame for the classification
        if The logo is real then
            Mark this shot as replay shot
        end if
    end if
end for
```

3.5 Conclusion

This chapter shows the concept of dominant color detection and the soccer video stream divide into small video shots using shot boundary detection in the pre-processing phase. Afterwards, different video shots assign into semantic descriptors (long, medium, close-up, and audience) based on the camera position and the video editing techniques in shot processing phase. Finally, in the replay detection phase, the proposed system applies two machine learning algorithms, namely; support vector machine (SVM) and artificial neural network (ANN), for emphasizing important segments with championship logo appearance.

Chapter 4

Event detection and summarization

4.1 Introduction

These three phases are described in detail in this chapter along with the steps involved and the characteristics feature for each phase, see the figure 4.1.

The objective of soccer video summarization is:

- Extract events or objects in the scene
- to produce general summaries for the most important moments in which TV viewers may be interested

The play field segmentation, events and objects detection play an important role in achieving the above described aims. Cinematography defines the rules, techniques, and conventions in film making. It specifies the most general and basic filming techniques, such as the rules for positioning the main objects on the screen and for setting the camera locations, as well as some tailored representations for specific domains, e.g., the use of scoreboard in sports. The video features that result from these generic and domain specific rules and conventions are referred to as cinematic features, and their efficient extraction and analysis makes fast and semantic video processing possible. one of the main disadvantages of all the cinematic features are that they aren't generic in the sense that they are specific for only one sport and dependent of the cinematic style

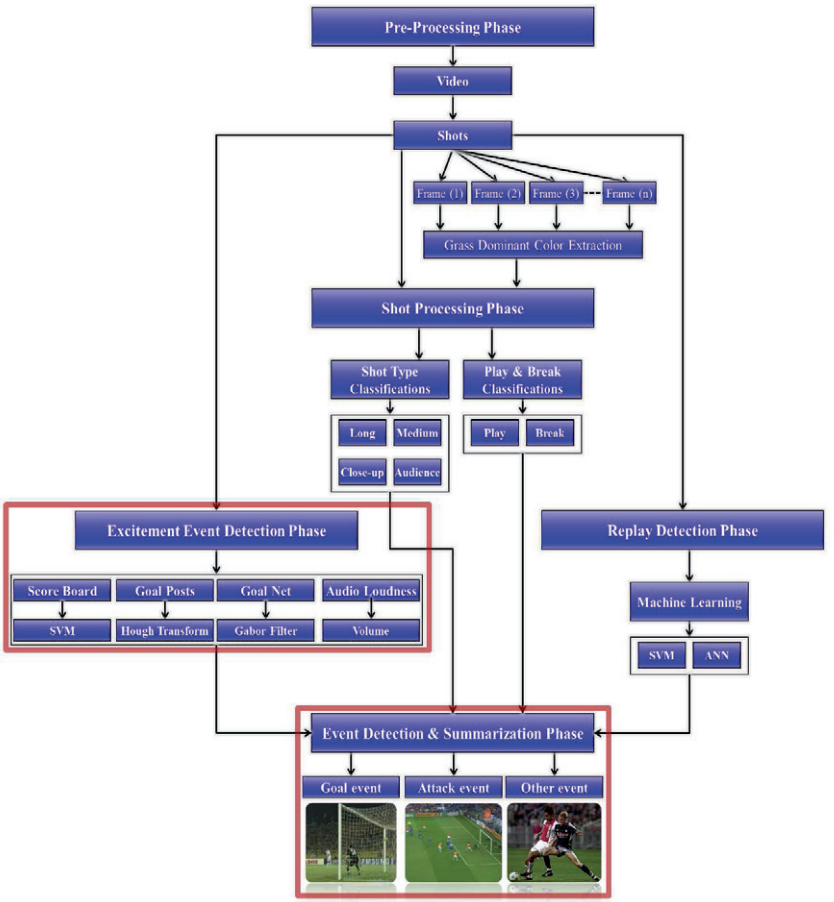


Figure 4.1: System architecture - event detection and summarization phases

employed by a particular broadcasting crew (1)(62).

4.2 Excitement event detection phase

This section presented the visual feature detection (cinematic features) algorithms for soccer video summarization. And we use all of those cinematic features to detect soccer events: Goal, attack and other events (fouls, injury and offside). The cinematic features will be used as a sequence, so we use the logo based replay to be the start of features sequence. Because of the robust detection of logo based replay with high recall and precision as you will see in the next chapter experimental results. Excitement event detection is based on three features as the following:

- **Scoreboard detection**
- **Goal mouth detection** by detecting both vertical goal posts and goal net.
- **Commentator loudness detection**

These three components are described in detail in the following subsections along with the steps involved and characteristics feature for each component.

4.2.1 Scoreboard detection

The scoreboard is a caption region distinguished from the surrounding region, which provides information about the score of the game or the status of the players (7)(2). The caption often appears at the bottom part of image frame for a short while and then disappears almost after appearing for 5 seconds. When the scoreboard is detected with enough confidence, it can undoubtedly provide the inference of goal event, because after every scored goal the scoreboard is displayed. Figure 4.2 shows different types of the existence of the scoreboard caption region. The lower third of each frame was checked for containing a scoreboard by applying algorithm (2) the same algorithm for logo based detection replay by training the support vector machine (SVM) and artificial neural network (ANN) with the training data about true and false samples of championships scoreboard.

Figure 4.2: The scoreboard caption region

4.2.2 Goal mouth detection

In soccer match video, we will observe more number of events during the importance activity. For example, during the goal, we will observe the close-up of the player/goalkeeper who has contributed, the close-up of referee, celebration of the players by gathering, logo based replays. The users of our system can choose which type of events they want to see in the content of summarization video by selecting from different event detection types (goal, attack and other events). For soccer match video, the goal mouth scenarios can be selected as the highlighted candidates, for the reason that most of the exciting events occurs in the goal mouth area such as (goal, shooting, penalty, direct free kick, .. etc), as shown in Figure 4.3: (a) and (b). On the other hand, the non-goal mouth scenarios often consist of the dull passes in the mid field, defense and offense or some other shots to the audiences or coaches, etc, which are not considered as exciting as the former, in Figure 4.3: (c) and (d). So we managed to extract goal mouth scenarios from the soccer video as the highlighted candidates. Figure 4.3 shows two scenarios which illustrate the differences between goal mouth shots and non-goal mouth shots (63). The proposed system detect the goal mouth by detecting both vertical goal posts and goal net.

4.2.2.1 Hough transform: an overview

The Hough transform (HT), named after Paul Hough who patented the method in 1962, is a powerful global method for detecting edges. It transforms between the Cartesian

Figure 4.3: A goal mouth appearance indicates potential exciting play a (a) and (b), but in (c) and (d) are non-goal mouth that illustrate non-exciting play

space and a parameter space in which a straight line (or other boundary formulation) can be defined.

In automated analysis of digital images, a subproblem often arises of detecting simple shapes, such as straight lines, circles or ellipses. In many cases an edge detector can be used as a pre-processing stage to obtain image points or image pixels that are on the desired curve in the image space. Due to imperfections in either the image data or the edge detector, however, there may be missing points or pixels on the desired curves as well as spatial deviations between the ideal line/circle/ellipse and the noisy edge points as they are obtained from the edge detector. For these reasons, it is often non-trivial to group the extracted edge features to an appropriate set of lines, circles or ellipses. The purpose of the Hough transform is to address this problem by making it possible to perform groupings of edge points into object candidates by performing an explicit voting procedure over a set of parameterized image objects (64).

The simplest case of Hough transform is the linear transform for detecting straight

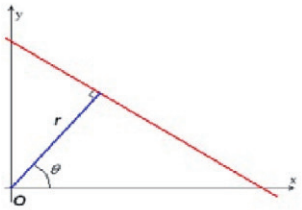


Figure 4.4: Hough transform for detecting lines

lines. In the image space, the straight line can be described as y = mx + b and can be graphically plotted for each pair of image points (x,y). In the Hough transform, a main idea is to consider the characteristics of the straight line not as image points x or y, but in terms of its parameters, here the slope parameter m and the intercept parameter b. Based on that fact, the straight line y = mx + b can be represented as a point (b, m) in the parameter space. However, one faces the problem that vertical lines give rise to unbounded values of the parameters m and b. For computational reasons, it is therefore better to use a different pair of parameters, denoted r and θ (theta), for the lines in the Hough transform. as you see the figure 4.4.

The parameter r represents the distance between the line and the origin, while θ is the angle of the vector from the origin to this closest point (see Coordinates as in figure 4.4). Using this parametrization, the equation of the line can be written as:

$$y = (\frac{\cos\theta}{\sin\theta})x + \frac{r}{\sin\theta} \tag{4.1}$$

which can be rearranged as the following equation:

$$r = x\cos\theta + y\sin\theta \tag{4.2}$$

It is therefore possible to associate to each line of the image a couple (r, θ) which is unique if $\theta \in [0,\pi]$ and r $\in$ R, or if $\theta \in [0,2\pi]$ and r $\geq$ 0. The (r, θ) plane is sometimes referred to as Hough space for the set of straight lines in two dimensions. This representation makes the Hough transform conceptually very close to the two dimensional

Radon transform. (They can be seen as different ways of looking at the same transform).

An infinite number of lines can pass through a single point of the plane. If that point has coordinates (x_0, y_0) in the image plane.

$$r(\theta) = x_0 \cdot \cos\theta + y_0 \cdot \sin\theta \tag{4.3}$$

where the parameter r represents the distance between the line and the origin.

This corresponds to a sinusoidal curve in the (r,θ) plane, which is unique to that point. If the curves corresponding to two points are superimposed, the location (in the Hough space) where they cross correspond to lines (in the original image space) that pass through both points. More generally, a set of points that form a straight line will produce sinusoids which cross at the parameters for that line. Thus, the problem of detecting collinear points can be converted to the problem of finding concurrent curves.

4.2.2.2 Vertical goal posts detection

The two vertical goal posts are distinctively characterized by their vertical strips of white and grow connected pixel gray values of white. we are primarily looking for a vertical strip of white as shown figure 4.5. The two vertical goal posts are distinctively characterized by their vertical strips of white and grow connected pixel gray values of white. Hough transform is used for detecting the two goal posts, as shown in figure 4.6. Algorithm (4) presents the steps applied to each frame for detecting the vertical goal posts (65).

4.2.2.3 Gabor filter: an overview

In image processing, a Gabor filter, named after Dennis Gabor, is a linear filter used for edge detection. Frequency and orientation representations of Gabor filters are similar to those of the human visual system, and they have been found to be particularly appropriate for texture representation and discrimination. In the spatial domain, a 2D Gabor filter is a Gaussian kernel function modulated by a sinusoidal plane wave. The Gabor filters are self similar: all filters can be generated from one mother wavelet by

Figure 4.5: Different view of vertical goal post

Figure 4.6: Hough transform detection for the vertical goal posts

Algorithm 4 Vertical Goal Posts Detection

1: **for** each frame **do**
2: Use K-means clustering to convert each frame to binary image using squared Euclidean distances measure
3: Given a set of observations $(x_1, x_2, , x_n)$, where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k sets $(k \leq n)$ $S = S_1, S_2, , S_k$ so as to minimize the within-cluster sum of squares (WCSS)

$$arg_S min \sum_{i=1}^{k} \sum_{x_j \in S_i} \| x_j - \mu_i \|^2 \tag{4.4}$$

where, μ_i is the mean of points in S_{i-1}
4: Use Hough transform to detect the two goal posts

$$rho = x * cos(\theta) + y * sin(\theta) \tag{4.5}$$

where, rho is the distance from the origin to the line along a vector perpendicular to the line, and θ is the angle between the x-axis and this vector
5: **if** The overlap between the vertical parallel lines greater than 80% **then**
6: mark this frame as goal post frame
7: **end if**
8: **end for**

dilation and rotation (66).

Its impulse response is defined by a harmonic function multiplied by a Gaussian function. Because of the multiplication convolution property (Convolution theorem), the Fourier transform of a Gabor filter's impulse response is the convolution of the Fourier transform of the harmonic function and the Fourier transform of the Gaussian function. The filter has a real and an imaginary component representing orthogonal directions. The two components may be formed into a complex number or used individually.

Complex:

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \exp\left(i\left(2\pi\frac{x'}{\lambda} + \psi\right)\right) \tag{4.6}$$

Real:

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \cos\left(2\pi \frac{x'}{\lambda} + \psi\right) \tag{4.7}$$

Imaginary:

$$g(x, y; \lambda, \theta, \psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \sin\left(2\pi \frac{x'}{\lambda} + \psi\right) \tag{4.8}$$

where:

$$x' = x \cos\theta + y \sin\theta \tag{4.9}$$

and

$$y' = -x \sin\theta + y \cos\theta \tag{4.10}$$

In this equation, λ represents the wavelength of the sinusoidal factor, θ represents the orientation of the normal to the parallel stripes of a Gabor function, ψ is the phase offset, σ is the sigma of the Gaussian envelope and γ is the spatial aspect ratio, and specifies the ellipticity of the support of the Gabor function.

4.2.2.4 Goal net detection

Detection of the two vertical goal posts isn't sufficient for possible exciting play. So, there still a need for an extra step to increase the accuracy of goal mouth appearances detection. Accordingly, the proposed system checks goal post frames for goal net existence using Gabor filter (7). The Gabor filter is used due to that the goal net has a unique pattern and repeated many times.

The Gabor filter is basically a Gaussian filter, with variances sx and sy along x and y-axes, respectively. the sx and sy are modulated by a complex sinusoid, with center frequencies U and V along x and y-axes, respectively. The Gabor filer is described by the following equations.

$$G = \exp((\frac{-1}{2}(\frac{\acute{x}}{s\acute{x}})^2 + (\frac{\acute{y}}{s\acute{y}})^2) * \cos 2 * \Pi * f * \acute{x}) \tag{4.11}$$

$$\acute{x} = x * \cos(\theta) + y * \sin(\theta); \tag{4.12}$$

$$\acute{y} = y * \cos(\theta) - x * \sin(\theta); \tag{4.13}$$

Where, sx and sy: variances along x and y-axes, respectively, f: frequency of the sinusoidal function, θ: the orientation of Gabor filter, and G: The output filter.

4.2.3 Commentator loudness detection

In this section, we search for a set of speech parameters that would be in some way correlated with the excitement level observed in commentators and, hence, would allow for an automatic speech based spotting of key moments in sports.

Loudness, silence and pitch generated by a commentator and/or crowd are effective measurements for detecting excitement. The volume is one the most frequently used and simplest audio features. As an indication of the loudness of sound, volume is very useful for soccer video analysis (50). The volume of each audio frame is calculated using equation (4.14):

$$Volume = \frac{1}{N} * \sum_{n=1}^{N} |x(n)| \tag{4.14}$$

Where N is the number of frames in a clip and $x(n)$ is the sample value of the nth frame.

To calculate pitch and silence, we applied the sub-harmonic toharmonic ratio based pitch determination in (67) for its reliability. Louder, less silence, and higher pitch audio frames are identified by using dynamic thresholds presented in (50). So, we can detect the excitement shots.

4.3 Event detection and summarization phase

When we are talking about soccer game, we can refer to a continuous sports which mean that if there is an existence of such a break during the match, it can an indicator of the occurrence of important event such as (goal, goal attempts shots, red/yellow

card, fouls and offside). Therefore the summarization process which we aimed for can be recognized by a combination of these events, for example the summarized segment may contains only the goal shots, goal attempts or penalty shots that can be described as important events (2)(5).

In this approach, we are going to highlight the most important events such as (Goals, attacks, etc), facilitating the process of automatic match, save the viewer's time, and introduce the technology of computer based summarization into sports field. The proposed system highlights the most important events during the soccer match into three classes (goal, attack, and the other event). Figure 4.7 shows the different event type classification.

Highlight detection cannot indicate how much interesting a highlight is, i.e., the confidence or important value. In this subsection, we present a scheme to evaluate the confidence of highlights. It is obvious that we have to rely on domain knowledge to give the evaluations of the highlights, for the highlights in different sports game represent different meaning. In soccer domain, the highlights usually occur at the following cases: a goal, a shoot, an interesting attack, severe foul (such as card), offside, and others (1)(2).

In this work, we evaluate a highlight with the following rules (sequence of features) for classify the events into goal, attack, and other events. Table 4.1 shows the cinematic features are used for each event detection.

4.3.1 Goal event detection

A goal is scored when the whole soccer ball passes the goal line between the goal posts and under the crossbar. However, it is difficult to verify these conditions automatically and reliably by the state of the art video processing algorithms. The occurrence of a goal event leads to a break in the game (2). Finally, the restart of the game is usually captured by a long shot view. During this break, the producers convey the emotions on the field to the TV audience and show one or more replay(s) for a better visual experience. The emotions of players are captured by one or more close-up views of the actors of the goal event, such as the scorer and the goalie, and by shots of the audience celebrating the goal. Furthermore, several slow motion replays of the goal event

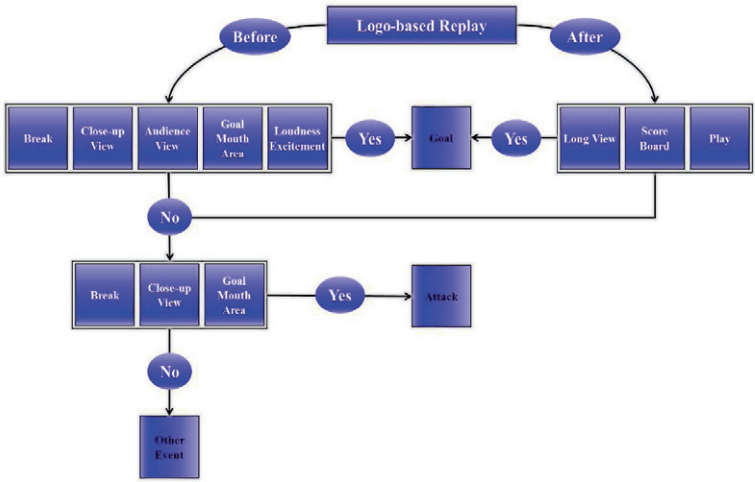


Figure 4.7: Event type classification

from different camera positions are shown. Finally, the restart of the game is usually captured by a long shot view. Figure 4.8 illustrates the instantiation of such a pattern of sequence of images for the first goal of Netherlands in Netherlands vs. Italy soccer match in euro 2008 (2).

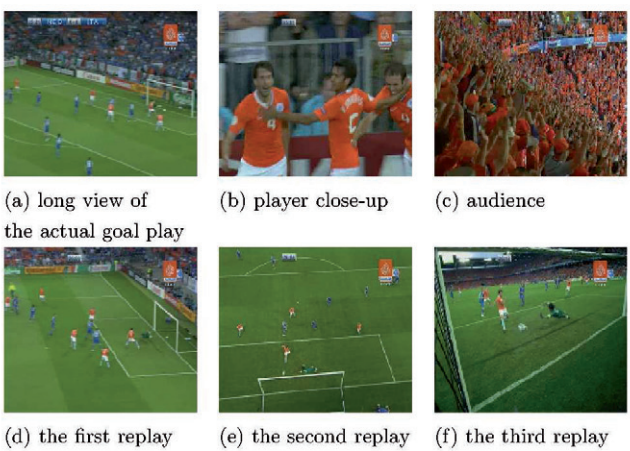
4.3.1.1 Goal event features

As you see in table 4.1, the following features used to detect the goal event that occurs between the long shot resulting in the goal event and the long shot view that shows the restart of the game:

- **Replay Duration (RD)** due to goal events the duration no less than 20 and no more than 60 seconds.
- **Goal Mouth (GM)**
- **Close-up View (CV)**

Feature	Description	Goal Event	Attack Event	Other Event
Replay Duration (RD)	Usually, the longer a replay scene is, the more attractive the segment is.	Yes	Yes	Yes
Goal Mouth (GM)	The goal mouth appearance are often shown before and within the replay scene in the cases of interesting goal, shoot and attack.	Yes	Yes	
Close-up View (CV)	This is a shot of a close-up of a player who scored the goal.	Yes	Yes	Yes
Audience View (AV)	An excited audience shot will be displayed after an interesting event according to general video cinematic feature.	Yes		
Long View (LV)	The long view to define the restart of the match after a replay (break).	Yes	Yes	Yes
Scoreboard Detection (SD)	This view are shown after a successful goal (a score). In these views, the caption containing score information is usually superimposed into long field-views.	Yes		
Commentator Loudness (CL)	Loudness, silence and pitch generated by a commentator are effective measurements for detecting excitement.	Yes		

Table 4.1: Event detection features



(a) long view of the actual goal play (b) player close-up (c) audience

(d) the first replay (e) the second replay (f) the third replay

Figure 4.8: An example of goal broadcast: the temporal order is from (a) to (f)

- **Audience View (AV)**
- **Long View (LV)**
- **Scoreboard Detection (SD)**
- **Commentator Loudness(CL)**

All these features used to indicate a goal event.

4.3.2 Attack and other event detection

Attack events may also match a lot of goal event features, although not as consistently as goals. The addition of attack events in the summaries may even be desirable since each of these events consists of interesting shots (68). There are other interesting events such as: fouls, cards, injure, or offside. The addition of these events in the summaries may even be desirable in order for each event to contain interesting shots. Therefore, more of users may enjoy watching interesting fouls and offside events.

4.3.2.1 Attack event features

As given in table 4.1, the following features are used to detect attack event that occurs between two pairs of replay logos:

- **Replay Duration (RD)** due to attack events the duration is no less than 15 and no more than 35 seconds.
- **Goal Mouth (GM)**
- **Close-up View (CV)**
- **Long View (LV)**

4.3.2.2 Other event features

Also as given in table 4.1, the following features are used to detect the other event that occurs between two pairs of replay logos:

- **Replay Duration (RD)** due to other events the duration is no less than no more than 15 seconds.
- **Close-up View (CV)**
- **Long View (LV)**

4.4 Conclusion

This chapter shows the objective of soccer video summarization, which is to produce general summaries for the most important moments in which TV viewers may be interested. The play field segmentation, events and objects detection play an important role in achieving the above described aims. Also the visual feature detection (cinematic features) algorithms for soccer video summarization are presented. All cinematic features to detect soccer events as: goal, attack and other events (fouls, injury and offside) are used. So, in the excitement event detection phase, the proposed system uses both ML algorithms for detecting the scoreboard which contains an information about the score of the game. The proposed system also uses k-means algorithm and Hough line transform for detecting vertical goal posts and Gabor filter for detecting goal net. Finally, in the event detection and summarization phase, the proposed system highlights the most important events during the match.

Chapter 5

Experimental result and analysis

The methodology we designed for the proposed automatic soccer video summarization system has been completely introduced in previous chapters. In this chapter, we discuss the experimental results from the tests data. We first describe the data set in Section 5.1. In Section 5.2, experimental results for all proposed system phases is presented.

5.1 Data sets

The proposed system was evaluated using soccer match videos from five Championships: World Cup Championship 2010 (2 full matches and 2 half match videos), Africa Championship League 2010 (3 full match videos), Africa Championship League 2008 (2 full match videos), European Championship League 2008 (1 full match video), and Euro 2008 (2 full match videos) as shown in the table 5.1. So, we have 10 complete soccer match videos and 2 half soccer match videos, it's total duration exceed 1200 minutes of soccer video game.

Championship	Names of the Teams	Note
World Cup 2010	Portugal VS Korea	Full Match
	Argentina VS Korea	Half Match
	Germany VS England	Half Match
	Germany VS Australia	Full Match
Africa League 2010	Ivory Coast VS Algeria	Full Match
	Egypt VS Cameroon	Full Match
	Egypt VS Algeria	Full Match
Africa League 2008	Tunisia VS Cameroon	Full Match
	Egypt VS Ivory Coast	Full Match
Euro 2008	Roma VS Real Madrid	Full Match
	Netherlands VS Italy	Full Match
European League 2008	Chelsea VS Derby	Full Match

Table 5.1: Soccer match videos championships data set

There are two categories for the quality of the used recorded videos:

- **High resolution** where frame size of 480 x 360 pixels.
- **Low resolution** where frame size of 352 x 240 pixels.

During experiments, both quality categories were covered. All soccer videos are in Audio Video Interleave (AVI) format with a frame rate of 30 fps and an audio track that is sampled at 44.1 kHz.

Two indicators; namely, ***recall*** and ***precision***, have been designed and calculated for each one of the proposed system phases in order to evaluate the performance of the proposed system and measure the resulted accuracy at each phase. Recall and precision ratios are the basic measures used in evaluating search strategies. Recall is the ratio

of the number of relevant records retrieved to the total number of relevant records in the database. Precision is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. both recall and precision are usually expressed as a percentage (69).

5.2 Experimental results

Results of shot boundary detection stage are shown in table 5.2. For a number of soccer match videos from the different championships with a total duration of 4hrs: 47min : 17sec, it has been obtained that 2081 shot boundaries have been correctly detected, while 194 shot boundaries have been failed to be correctly detected. These 194 shot boundaries have been divided into 135 falsely detected shot boundaries with addition to 59 missed shot boundaries. Accordingly, recall ratio for shot boundary detection was 97.2%, whereas precision ratio was 93.9%.

Table 5.2: Shot boundary detection results

Duration (hh:mm:ss)	Correct	False	Miss	Recall	Precision
4:47:17	2081	135	59	97.2 %	93.9 %

Table 5.3 and figure 5.1 shows the results of shot views classification stage. The proposed system resulted 925, 97, 46, 15 shot views classified as long, medium, close-up and audience shot views, respectively, out of the total provided *long* shot views to the system. Also, the proposed system resulted 83, 807, 17, 19 shot views classified as long, medium, close-up and audience shot views, respectively, out of the total provided *medium* shot views to the system. Moreover, the proposed system resulted 0, 43, 450, 25 shot views classified as long, medium, close-up and audience shot views, respectively, out of the total provided *close-up* shot views to the system. Furthermore, the proposed system resulted 0, 3, 12, 87 shot views classified as long, medium, close-up and audience shot views, respectively, out of the total provided *audience* shot views to the system. Recall ratios for long, medium, close-up and audience shot views classification have been obtained to be 91.8%, 87.1%, 85.7%, and 59.6%, respectively. Precision ratios for long, medium, close-up and audience shot views classification have been obtained to be

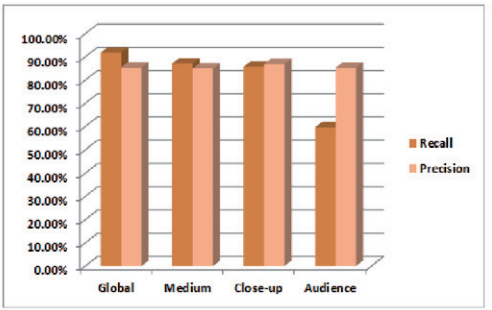


Figure 5.1: Shot classification results

85.4%, 84.9%, 86.9%, and 85.3%, respectively.

Table 5.3: Shot classification results

Shot-Type	Long	Medium	Close-up	Audience
Long	925	97	46	15
Medium	83	807	17	19
Close-up	0	43	450	25
Audience	0	3	12	87
Recall	91.8%	87.1%	85.7%	59.6%
Precision	85.4%	84.9%	86.9%	85.3%

Table 5.4 and figure 5.2 illustrates results of both SVM based and ANN based logo replay detection stage. Compared to the performance results obtained using SVM classifier, the proposed system attained good ANN based performance results concerning recall ratio, however it attained poor ANN based performance results concerning precision ratio.

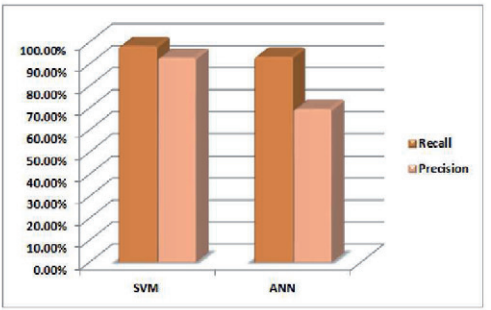


Figure 5.2: Evaluation of logo based replay using SVM and ANN

Table 5.4: Evaluation of logo based replay using SVM and ANN

Factors	SVM	ANN
Duration (hh:mm:ss)	1:53:39	1:53:39
Correct	103	98
False	8	43
Miss	2	7
Recall	98.1%	93.3%
Precision	92.8%	69.5 %

Table 5.5 and table 5.6 show the results of score board and goal mouth detection, respectively. For scoreboard detection, SVM classifier has been used whereas both Gabor filter and Hough transform have been used for goal mouth detection.

Table 5.5: Evaluation of scoreboard detection

Duration (hh:mm:ss)	Correct	False	Miss	Recall	Precision
1:53:39	68	5	1	98.5%	93.1 %

Table 5.6: Evaluation of goal mouth detection

Duration (hh:mm:ss)	Correct	False	Miss	Recall	Precision
1:30:42	247	25	11	95.7%	90.8%

Table 5.7 and figure 5.3 shows the confusion matrix for event detection and summarization resulted from the proposed system.

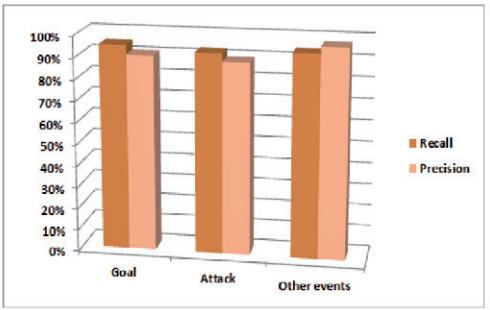


Figure 5.3: Results of event detection and summarization

Table 5.7: Confusion matrix for event detection and summarization

Event Detection	Goal	Attack	Other events
Goal	57	3	0
Attack	6	176	8
Other events	0	18	283
Recall	95%	92.6%	94%
Precision	90.5%	89%	97.3%

Chapter 6

Conclusions and future work

6.1 Conclusions

In this thesis, a novel mid level representation for soccer video parsing was introduced for bridging the gap between low level features such as color, motion and texture and semantic meanings. The machine learning based system proposed in this thesis for broadcast soccer videos summarization is composed of five phases; namely, pre-processing phase, shot processing phase, replay detection phase, excitement event detection phase, and event detection and summarization phase.

Pre-processing phase is to segment the whole video stream into small video shots. By detecting the dominant color in the video frame, then the shot boundary detection algorithm is applied in order to output video shots based on dominant color derived features. The dominant color is the color that fills most of the given area, and it is different among various play fields. In this thesis, we are concerned only with the soccer game, which has a green color for the playing field. Shot boundary detection is separated views that come from multiple cameras positioned at different locations. It can be realized that while changing from one camera to another, this indicates a start of a new shot and marks a boundary of a new shot.

Shot processing phase is converting a soccer video stream into a semantic descriptor sequence. Each shot can be classified into categories (close-up, audience, long, medium, play, and break) in the shot processing phase. This phase applies two types of classi-

fication; namely, shot type classification and play / break classification, to the video shots resulted from the pre-processing phase. Cinematographers classify a shot into one of four categories long, medium, close-up and audience (out of field) shot classes. The sports video is a repetitions of play and break scenes. Some researchers claim that play sequences in sports videos are self consumable because most users naturally focus their attention on events that happen within plays.

Replay detection phase is a good semantic cue to detect events due to frequently happening after every event in sports games, especially in the soccer domain. Therefore, we use Replay as one basic event in our method. The proposed system use logo based replay detection as identification of a wide variety of logos used by different broadcasters. For detecting the logos, the proposed system applies two machine learning algorithms, namely; support vector machine (SVM) and artificial neural network (ANN).

Excitement event detection phase is presented by the visual feature detection (cinematic features) algorithms for soccer video summarization. We use all of those cinematic features to detect soccer events as: goal, attack and other events (fouls, injury and offside). The cinematic features will be used as a sequence, so we use the logo based replay to be the start of features sequence. The scoreboard is a caption region distinguished from the surrounding region, which provides information about the score of the game or the status of the players. For soccer match video, the goal mouth scenarios can be selected as the highlighted candidates, for the reason that most of the exciting events occurs in the goal mouth area such as (goal, shooting, penalty, direct free kick, .. etc). The proposed system detect the goal mouth by detecting both vertical goal posts and goal net. Loudness, silence and pitch generated by a commentator and/or crowd are effective measurements for detecting excitement.

Event detection and summarization phase highlight the most important events such as (Goals, attacks, etc), facilitating the process of automatic match, save the viewer's time, and introduce the technology of computer based summarization into sports field. The proposed system highlights the most important events during the soccer match

into three classes (goal, attack, and the other event).

The proposed system was evaluated using soccer match videos from five international soccer championships: World Cup Championship 2010, Africa Championship League 2010, Africa Championship League 2008, European Championship League 2008, and Euro 2008. So, we have 10 complete soccer match videos and 2 half soccer match videos, it's total duration exceed 1200 minutes of soccer video game. Compared to the performance results obtained using SVM classifier, the proposed system attained good ANN based performance results concerning recall ratio, however it attained poor ANN based performance results concerning precision ratio. Accordingly, it has been concluded that using the SVM classifier is more appropriate for soccer videos summarization than ANN classifier as seeing the experimental results in the previous chapter.

6.2 Future work

For future research, we can increase the number of soccer videos and championships being examined in order to get more accurate results. Moreover, different new machine learning techniques may be applied.

Several areas that may be promising for future research is described in the following:

- How to further improve the performances for both the approaches by using more generic features for the unit based approach and by using more domain knowledge and features for the frame based approach.
- The computation of the initial dominant color statistics was based on the ratio of dominant color pixels in the training set that was input by a human operator. Although this did not pose any problem for the applications because it can be completed before the start of the game, automatic computation of thresholds for dominant color region detection should be considered for our future study.
- Proposed a novel trajectory based algorithm for automatically detecting and tracking the ball in broadcast soccer video. We aim to extend our method us-

ing motion trajectories and shapes as low level evidence in addition to color and texture.

- The research on finding relationships between audio and video features for soccer video analysis is a promising avenue. More research is based on the loudness, silence and pitch generated by a commentator and/or crowd are effective measurements for detecting excitement.

- Extension of the proposed approach to different sports, such as American football, basketball, and baseball, which require different event and object detection modules, will be addressed in the future.

- For future research, we can increase the number of soccer match videos, international championships, local league, and data set being examined.

References

[1] E. Lotfi and H.R. Pourreza. **Event Detection and Automatic Summarization in Soccer Video.** *4th Iranian Conference on Machine Vision and Image Processing (MVIP07), Mashhad, Iran*, 2007. 1, 21, 32, 43, 52

[2] A. Ekin; A.M. Tekalp and R. Mehrotra. **Automatic Soccer Video Analysis and Summarization.** *IEEE Transactions on Image processing, Vol. 12, No. 7, pp. 796-807*, 2003. 4, 15, 16, 23, 24, 26, 28, 29, 30, 43, 52, 53

[3] H. Pan; P. Beek and M. Sezan. **Detection of slow-motion replay segments in sports video for highlights generation.** *The IEEE International Conference on Acoustics, Speech, Signal Processing (ICASSP'01), Salt Lake City, US, Vol. 3, pp. 1649-1652*, 2001. 4

[4] B. Li H. Pan and M. Sezan. **Automatic detection of replay segments in broadcast sports programs by detection of logos in scene transitions.** *The IEEE International Conference on Acoustics, Speech, Signal Processing (ICASSP'02), Orlando, Florida, USA, pp. 3385-3388*, 2002. 4

[5] Hossam M. Zawbaa; Nashwa El-Bendary; Aboul Ella Hassanien; Gerald Schaefer and Sang-Soo Yeo. **Support Vector Machine based Logo Detection in Broadcast Soccer Videos.** *The 16th Online World Conference on Soft Computing in Industrial Applications (WSC16) Advances in Intelligent and Soft Computing, Springer*, 2011. 6, 28, 31, 32, 52

[6] Hossam M. Zawbaa; Nashwa El-Bendary; Aboul Ella Hassanien and Tai hoon Kim. **Machine Learning-based Soccer Video Summarization**

System. *International Conference on Multimedia, Computer Graphics and Broadcasting (MulGraB 2011), CCIS/LNCS series Springer, (Indexed by SCOPUS, EI), Korea, pp. 19-28*, 2011. 6

[7] HOSSAM M. ZAWBAA; NASHWA EL-BENDARY; ABOUL ELLA HASSANIEN AND AJITH ABRAHAM. **SVM-based Soccer Video Summarization System**. *The Third IEEE World Congress on Nature and Biologically Inspired Computing (NaBIC 2011) Salamanca University, Spain, pp. 7-11*, 2011. 6, 43, 50

[8] C.G.M. SNOEK AND M. WORRING. **A State-of-the-art Review on Multimodal Video Indexing**. *Proc. of the 8th Annual Conference of the Advanced School for Computing and Imaging, Lochem, Netherlands, pp. 194-202*, 2002. 11

[9] C.W. NGO; T.C. PONG AND H. J. ZHANG. **Recent Advances in Content based Video Analysis**. *International Journal of Image and Graphics, Vol. 1, No. 3, pp. 445-468*, 2001. 11, 12

[10] G.AHANGER AND T. D. C. LITTLE. **A Survey of Technologies for Parsing and Indexing Digital Video**. *International Journal of Visual Communication and Image Representation, Vol. 7, No. 1, pp. 28-43*, 1996. 11

[11] R. BRUNELLI; O. MICH AND C. M. MODENA. **A Survey on the Automatic Indexing of Video Data**. *International Journal on Visual Communication and Image Representation, Vol. 10, pp. 78-112*, 2001. 11

[12] C. W. NGO; T. C. PONG AND R. T. CHIN. **A Robust Wipe Detection Algorithm**. *Asian Conference on Computer Vision, Vol. 1, pp. 246-251*, 2000. 11, 22, 29

[13] M. S. DREW; S. N. LI AND X. ZHONG. **Video Dissolve and Wipe Detection via spatio - temporal Images of Chromatic Histogram Differences**. *International Conference on Image Processing, Vol. 3, pp. 929-932*, 2000. 11

[14] N. VASCONCELOS AND A. LIPPMAN. **Statistical Models of Video Structure for Content Analysis and Characterization**. *IEEE Transactions on Image Processing, Vol. 9, No. 1, pp. 3-19*, 2000. 11

[15] P. Bouthemy; M. Gelgon and F. Ganansia. **A Unified Approach to Shot Change Detection and Camera Motion Characterization.** *IEEE Transactions on Circuits and Systems for Video Technology, Vol. 9, No. 7, pp. 1030-1044*, 1999. 11

[16] R.A. Joyce and B. Liu. **Temporal Segmentation of Video Using Frame and Histogram Space.** *IEEE Transactions on Multimedia, pp. 130-140*, 2006. 11

[17] R. Lienhart. **Comparison of Automatic Shout Boundary Detection Algorithms.** *SPIE Conference on Storage and Retrieval for Image and Video Databases, pp. 290-301*, 1999. 11

[18] S. W. Lee; Y. M. Kim and S. W. Choi. **Fast Scene Change Detection Using Direct Feature Extraction from MPEG Compressed Videos.** *IEEE Transactions on Multimedia, pp. 240-254*, 2000. 11

[19] L. F. Cheong. **Scene-based Shot Change Detection and Comparative Evaluation.** *International Journal on Computer Vision and Image Understanding, Vol. 79, No. 2, pp. 224-235*, 2000. 11

[20] U. Gargi; R. Kasturi and S.H. Strayer. **Performance Characterization of Videoshot change Detection Methods.** *IEEE Transactions on Circuits and Systems for Video Technology, Vol. 10, No. 1, pp. 1-13*, 2000. 11

[21] B. L. Yeo and B. Liu. **Unified Approach to Temporal Segmentation of Motion JPEG and MPEG video.** *International Conference on Multimedia Computing and systems, pp. 2-13*, 1995. 11

[22] H. J. Zhang; A. Kankanhalli and S. W. Smoliar. **Automatic Partitioning of Fullmotion Video.** *Multimedia Systems Journal, Vol. 1, No. 1, pp. 10-28*, 1993. 11

[23] J. Meng; Y. Juan and S. F. Chang. **Scene Change Detection in an MPEG Compressed Video Sequence.** *IST/SPIE Symposium Proceedings, Vol. 2419, No. 1, pp. 14-25*, 1995. 11

[24] V. KOBLA; D. DEMENTHON AND D. DOERMANN. **Special Effect Edit Detection Using VideoTrails: a Comparison with Existing Techniques.** *SPIE Conference on Storage and Retrieval for Image and Video Databases VII, pp. 302-310*, 1999. 11

[25] J. S. BORECZKY AND L. A. ROWE. **Comparison of Video Shot Boundary Detection Techniques.** *SPIE Conference on Storage and Retrieval for Image and Video Databases IV, pp. 170-179*, 1996. 11

[26] R. ZABIH; J. MILLER AND K. MAI. **A Feature-based Algorithm for Detecting and Classifying Scene Breaks.** *ACM Multimedia 95, pp. 189-200*, 1995. 11

[27] S. DAGTAS; W. A1-KHATIB; A. GHAFOOR AND R.L. KASHYAP. **Models for Motion-based Video Indexing and Retrieval.** *IEEE Transactions on Image Processing, Vol. 9, No. 1, pp.88-101*, 2000. 12

[28] S. F. CHANG; W. CHEN; H. J. MENG; H. SUNDARAM AND D. ZHONG. **A fully Automatic Content-based Video Search Engine Supporting Multiobject spatio-temporal Queries.** *IEEE Transactions on Circuits and Systems for Video Technology, Vol. 8, No. 5, pp. 602615*, 1998. 12

[29] Y. WU; Y. ZHUANG AND Y. PAN. **Content-based Video Similarity Model.** *ACM Multimedia 95, pp. 189-200*, 2000. 12

[30] W. CHEN AND S. CHANG. **Generating Semantic Visual Templates for Video Databases.** *IEEE International Conference on Multimedia Expo, vol. 3, pp. 13371340*, 2000. 12, 13

[31] B. CLARKSON AND A. PENTLAND. **Unsupervised Clustering of Ambulatory Audio and Video.** *International Conference On Acoustics, Speech, and Signal Processing, Vol. 6, pp. 3037 -3040*, 1999. 12, 13

[32] E. SAHOURIA AND A. ZAKHOR. **Content Analysis of Video Using Principal Components.** *IEEE Transactions on Circuits and Systems for Video Technology, Vol. 9, No. 8, pp. 1290-1298*, 1999. 12, 13

[33] W. Zhou; A. Vellaika and C. C. Jay Kuo. **Rule-based Video Classification System for Basketball Video Indexing**. *International workshop on Multimedia Information Retrieval, pp. 213-216*, 2000. 12, 13

[34] C. Kim and J.N. Hwang. **Fast and Robust Moving Object Segmentation in Video Sequences**. *IEEE international Conference on Image Processing, Kobe, Japan, pp. 131-134*, 1999. 13

[35] C. Kim and J-N Hwang. **Object-Based Video Abstraction and an Integrated Scheme for Online Processing**. *IEEE Transactions on Circuits and Systems for Video Technology (CSVT), pp. 303-311*, 2000. 13

[36] C. Kim and J. N. Hwang. **Object-Based Video Abstraction Using Cluster Analysis**. *IEEE International Conference on Image Processing, Greece, pp. 657-660*, 2001. 13

[37] M.R. Naphade and T.S. Huang. **Extracting Semantics from Audiovisual Content: The Final Frontier in Multimedia Retrieval**. *IEEE Transactions on Neural Networks, Vol.13, No. 4, pp. 793-810*, 2002. 13

[38] S. F. Chang; W. Chen and H. Sundaram. **Semantic Visual Templates Linking Features to Semantics**. *IEEE International Conference on Image Processing, Vol. 3, Chicago, IL, pp. 531535*, 1998. 13

[39] J. N Hwang and Y. Luo. **Automatic Object based Video Analysis and Interpretation: A Step toward systematic video understanding**. *invited special session talk in ICASSP, Orlando FL, pp. 4084-4087*, 2002. 13

[40] Y. H. Gong; L. T. Sin; C. H. Chuan; H. J. Zhang and M. Sakauchi. **Automatic Parsing of TV Soccer Programs**. *IEEE International Conference on Multimedia Computing and Systems, pp. 167-174*, 1995. 14

[41] D. Yow; B. L. Yeo; M. Yeung and B. Liu. **Analysis and Presentation of Soccer HighLights from Digital Video**. *Asian Conference on Computer Vision, pp. 499-503*, 1995. 14

[42] V. Tovinkere and R. J. Qian. **Detecting Semantic Events in Soccer Games: Towards a Complete Solution.** *IEEE International Conference on Multimedia and Expo, pp. 833-836*, 2001. 14

[43] O. Utsumi; K. Miura; I. Ide; S. Sakai and H. Tanaka. **An Object Detection Method for Describing Soccer Game from Video.** *IEEE International Conference on Multimedia and Expo, Vol. 1, pp. 45-48*, 2002. 15

[44] P. Xu et al. **Algorithms and Systems for Segmentation and Structure Analysis in Soccer Video.** *IEEE International Conference on Multimedia and Expo, Tokyo, Japan, pp. 721-724*, 2001. 15, 16, 24

[45] L. Y. Duan; M. Xu; X. D. Yu and Q. Tian. **A Unified Framework for Semantic Shot Classification in Sports Video.** *ACM, Juan- les-Pins, France, Vol. 7, No. 6, pp. 1066-1083*, 2002. 16

[46] H. J. Zhang; C. Y. Low; S. W. Smoliar and J. H. Wu. **Video Parsing, Retrieval and Browsing: An Integrated and Content-based Solution.** *ACM Multimedia'95, pp. 15-24*, 1995. 16, 27

[47] H. w. Kim and K. S. Hong. **Soccer Video Mosaicing using Self-Calibration and Line Tracking.** *International Conference on Pattern Recognition, Barcelona, Spain, pp.592-595*, 2000. 16

[48] T. Qing; L. Joo-Hwee; J. S. Jesse; S. Haiping and Qi Tian. **A generic mid-level representation for semantic video analysis.** *International conference on image processing, pp. 629-632*, 2004. 17

[49] C. W. Ngo; T. C. Pong and R. T. Chin. **Detection of Gradual Transitions through Temporal Slice Analysis.** *IEEE Conference on Computer Vision and Pattern Recognition, pp. 1036-1041*, 1999. 27

[50] D. Tjondronegoro; Y.P. Chen and B. Pham. **Sports video summarization using highlights and play-breaks.** *ACM SIGMM International Workshop on Workshop on Multimedia Information Retrieval (ACM MIR'03), Berkeley, USA, pp. 201-208*, 2003. 31, 51

[51] D. TJONDRONEGORO; Y.P. CHEN AND B. PHAM. **The power of play-break for automatic detection and browsing of self-consumable sport video highlights.** *In Multimedia Information Retrieval, pp. 267-274*, 2004. 31

[52] S. XING-HUA AND Y. JING-YU. **Inference and retrieval of soccer event.** *Journal of Communication and Computer, Vol. 4, No. 3*, 2007. 31

[53] V. VAPNIK. **Statistical Learning Theory.** *IEEE Transactions on Neural Networks, Vol. 10, No. 5, pp. 988-999*, 1999. 34, 35

[54] C.J. BURGES. **A tutorial on support vector machines for pattern recognition.** *Data Mining and Knowledge Discovery, Vol.2, No.2, pp.121-167*, 1998. 34

[55] M.R. ISLAM; M.U. CHOWDHURY AND W. ZHOU. **An innovative spam filtering model based on support vector machine.** *International Conference on Computational Intelligence for Modelling, Control and Automation, Vol.2, pp. 348-353*, 2005. 35, 36

[56] C.M. BISHOP. **Neural Networks for Pattern Recognition.** *Oxford University Press*, 1995. 37

[57] S. GROSSBERG. **Adaptive pattern classification and universal recoding: Parallel development and coding of neural feature detectors.** *Biological Cybernetics, Vol.23, pp. 121-134*, 1976. 37

[58] G. CARPENTER AND S. GROSSBERG. **Adaptive resonance theory.** *The Handbook of Brain Theory and Neural Networks. pp. 79-82*, 1995. 38

[59] B. YU AND D. ZHU. **Automatic thesaurus construction for spam filtering using revised back propagation neural network Expert Systems with Applications.** *Expert System and Applications, Tarrytown, NY, USA, pp. 18-23*, 2010. 38

[60] A. EKIN; A. M. TEKALP AND R. MEHROTRA. **Robust dominant color region detection with applications to sports video analysis.** *In Proceedings of IEEE ICIP, Vol. 1, pp. 21-24*, 2003. 39

[61] Bo Han; Yan Yan; Zhenghua Chen; Chang Liu and Weiguo Wu. **A general framework for automatic on-line replay detection in sports video.** *In Proceedings of the 17th ACM international conference on Multimedia (MM '09). ACM, New York, NY, USA, pp. 501-504*, 2009. 39

[62] M. Bertini; A. Del Bimbo; C. Torniai; C. Grana; R. Vezzani and R. Cucchiara. **Sports Video Annotation Using Enhanced HSV Histograms in Multimedia Ontologies.** *In Proceedings of the 14th International Conference of Image Analysis and Processing - Workshops (ICIAPW '07). IEEE Computer Society, Washington, DC, USA, pp. 160-170*, 2007. 43

[63] Zhao Zhao; Shuqiang Jiang; Qingming Huang and Qixiang Ye. **Highlight Summarization in Soccer Video Based on Goalmouth Detection.** *Asia-Pacific Workshop on Visual Information Processing, pp. 1-4*, 2006. 44

[64] C. Galambos; J. Kittler and J. Matas. **Progressive probabilistic hough transform for line detection.** *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1, pp. 554-560*, 1999. 45

[65] Kongwah Wan; Xin Yan; Xinguo Yu and Changsheng Xu. **Real-time goal-mouth detection in MPEG soccer video.** *In Proceedings of the eleventh ACM international conference on Multimedia (MULTIMEDIA '03). ACM, New York, NY, USA, pp. 311-314*, 2003. 47

[66] V. Kyrki; J.-K. Kamarainen and H. Kalviainen. **Simple Gabor feature space for invariant object recognition.** *Pattern Recognition Letters, pp. 311318*, 2004. 49

[67] X. Sun. **Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio.** *The IEEE International Conference on Acoustics, Speech, Signal Processing (ICASSP'02), Orlando, Florida, USA, Vol. 1, pp. 333-336*, 2002. 51

[68] A. Ekin. **Sports Video Processing for Description, Summarization and Search.** *PhD Thesis, University of Rochester*, 2003. 55

[69] F. Ye; J. Liu; B. Liu and K. Chai. **Duplicate page detection algorithm based on the field characteristic clustering**. *In ICWL Workshops, pp. 75-84*, 2010. 59

Declaration

I herewith declare that I have produced this thesis without the prohibited assistance of third parties and without making use of aids other than those specified; notions taken over directly or indirectly from other sources have been identified as such. This thesis has not previously been presented in identical or similar form to any other Egyptian or foreign examination board.

The thesis work was conducted from Hossam Mohammed Zawbaa Ismail to Cairo University under the supervision of Prof. Dr. Aboul Ella Hassanien, Dr. Nashwa El-Bendary, and Dr. Iman Atef El-Azab at faculty of computer science and information technology, Cairo University.

Cairo, Egypt
January, 2012